

Fenchel–Young Duality Gaps: Certified Early Stopping for Regularized Inverse Problems

Pierre-Cyril Aubin-Frankowski¹ and Yohann De Castro^{1,2,3}

¹*CERMICS, CNRS, ENPC, Institut Polytechnique de Paris, Marne-la-Vallée, France.*

²*Institut Camille Jordan, École Centrale Lyon, CNRS UMR 5208, France.*

³*Institut Universitaire de France (IUF).*

September 2026

Abstract

We study computable error bounds and certified early stopping for regularized inverse problems, where a data-fidelity term is traded against a regularizer. The analysis relies on an exact duality-gap identity that splits the total gap of $F(\Phi\mu) + \lambda R(\mu)$ into a data-fidelity Fenchel–Young loss and a regularizer Fenchel–Young loss,

$$\Delta(\mu, h) = L_F(\Phi\mu \parallel h) + \lambda L_R(\mu \parallel \eta), \quad \eta = -\Phi^*h/\lambda,$$

valid for any primal point μ and any dual point h . The data-fidelity term F is strictly convex, so wherever F^* is differentiable the loss $L_F(\Phi\mu \parallel h)$ is the Bregman divergence of F between the prediction $\Phi\mu$ and $\nabla F^*(h)$, and it vanishes exactly at *Mirror Alignment* $h = \nabla F(\Phi\mu)$. Evaluated at a dual-feasible point $\tilde{h}$, the gap $\Delta(\mu, \tilde{h})$ is computable and *oracle-free*, meaning that it uses no knowledge of the solution, and it bounds the suboptimality of μ . Under the source condition, the same Fenchel–Young losses give *a priori* bounds on the estimation and prediction errors. Their scale is the irreducible model and noise error $L_F(\Phi\mu^* \parallel h^*)$, which vanishes exactly when Mirror Alignment holds at the certificate. This gives an early-stopping rule: run the algorithm until the regularizer Fenchel–Young loss falls below a tolerance ϵ . A constructive version of the Brøndsted–Rockafellar theorem then turns the current pair into an exact *dual-feasible* one, and this proxy lifts to an exact primal certificate. We build the proxy by a proximal step in the geometry of the fidelity, with Bregman kernel F^* and tilted by the prediction $\Phi\mu$: it recovers the Euclidean step of Carlier when F is the squared error, and it reduces the duality gap by the regularizer Fenchel–Young loss, up to a second-order remainder that vanishes in the quadratic case. Our running example is the Generalized Beurling–Lasso (GBL), where R is the total-variation norm on signed measures. It contains the classical Beurling–Lasso, obtained with the squared error, and also covers robust, logistic, entropic and inverse-optimal-transport losses. The same duality gap certifies deep-learning optimizers such as Lion-K and Muon, in their proximal form, as solvers of the regularized program. A companion paper by the same authors builds on these error bounds to establish exact support recovery for the GBL under a non-degenerate source condition.

Keywords. Fenchel–Young losses; duality gap; certified early stopping; Brøndsted–Rockafellar theorem; total-variation regularization; Beurling–Lasso; Mirror Alignment; Bregman divergence.

MSC 2020. 90C25, 49N15, 65K10, 94A12, 49M29.

1 Introduction

Recovering structured signals from indirect, noisy measurements is a classical problem in inverse problems and signal processing. We focus on convex programs that combine a strictly convex *data-fidelity* term $F : \mathcal{H} \rightarrow \mathbb{R}$, acting on the prediction $\Phi\mu$ in a separable Hilbert space $\mathcal{H}$, with a lower semi-continuous proper convex *regularizer* R , acting on a Banach-dual variable μ ,

$$\min_{\mu} \left\{ F(\Phi\mu) + \lambda R(\mu) \right\}, \quad (1.1, F+R)$$

where Φ is a bounded weak-* continuous forward operator and $\lambda > 0$ trades fidelity against regularization. All the results of this paper (duality-gap decomposition, Fenchel–Young error bounds, Mirror Alignment, Brøndsted–Rockafellar early stopping, and its mirror-prox refinement) are developed at this level of generality; specific choices of R (the total-variation norm of measures, group/joint TV, the nuclear norm, analysis sparsity, ...) instantiate the framework on the application domains catalogued in Section 4.

A canonical instance, used as our running example throughout the paper, is the *Generalized Beurling–Lasso* (GBL), obtained by specializing (1.1, F+R) to signed measures $\mu \in \mathcal{M}(\mathcal{X})$ on a closed domain $\mathcal{X} \subseteq \mathbb{R}^d$, possibly unbounded, with the total-variation norm $R(\mu) = \|\mu\|_{\text{TV}}$:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ F(\Phi\mu) + \lambda \|\mu\|_{\text{TV}} \right\}. \quad (1.2, \text{GBL})$$

The classical *Beurling–Lasso* (BLASSO) of Bredies and Pikkarainen [2013], De Castro and Gamboa [2012] is the further specialization of GBL to the squared-error data fidelity $F(h) = \frac{1}{2} \|h - y\|_{\mathcal{H}}^2$ with measurement $y \in \mathcal{H}$:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ \frac{1}{2} \|\Phi\mu - y\|_{\mathcal{H}}^2 + \lambda \|\mu\|_{\text{TV}} \right\}. \quad (1.3, \text{BLASSO})$$

Moving from BLASSO to the GBL form (1.2, GBL), and then to the abstract F+R program (1.1, F+R), is needed to model settings where additive Gaussian noise is not appropriate, such as photon-limited imaging which uses Poisson regression; binary classification which uses logistic losses; robust regression which uses losses that tolerate heavy-tailed noise, such as the Huber loss; or inverse optimal transport which uses an entropic generating functional. The same Fenchel–Young geometry applies to all of these cases, and they are treated in this paper.

1.1 Contributions

We organize our contributions as answers to two questions about the convex program (1.1, F+R) and its GBL specialization (1.2, GBL). Our main message is that algebraic manipulations on Fenchel–Young gaps/losses are very informative and provide early stopping certificates.

Q1. When does the duality gap give a computable error bound on the optimization error? We prove that for any primal variable μ and any dual variable h , the total duality gap admits an exact algebraic decomposition (Theorem 1):

$$\Delta(\mu, h) = L_F(\Phi\mu \parallel h) + \lambda L_R(\mu \parallel \eta), \quad \text{where} \quad \eta = -\frac{\Phi^*h}{\lambda}.$$

Here L_F is the data-fidelity Fenchel–Young loss (equal, for any $h \in \text{int dom } F^*$, to the Bregman divergence of F between $\Phi\mu$ and $\nabla F^*(h)$, and vanishing exactly at *Mirror Alignment* $h = \nabla F(\Phi\mu)$), and L_R penalizes the violation of the subgradient condition for the regularizer. A primal–dual pair (μ, h) meeting both the source condition $\eta \in \partial R(\mu)$ and Mirror Alignment $h = \nabla F(\Phi\mu)$ is optimal, and conversely every primal solution μ admits such a dual h (Theorem 2); at optimality, primal prediction and dual variable are Legendre–Fenchel conjugates. The loss L_F then gives two guarantees: (i) evaluated at a dual-feasible proxy $\tilde{h}$, the duality gap $\Delta(\mu, \tilde{h})$ is a *computable*, oracle-free upper bound on the objective suboptimality (Theorem 1); and (ii) under the source condition, the same Fenchel–Young losses bound the estimation and prediction errors *a priori* through the scalar $L_F(\Phi\mu^* \parallel h^*)$ corresponding to the irreducible model/noise error (Theorem 3).

Q2. How early can an algorithm be stopped while still certifying admissibility? Practical algorithms drive the regularizer Fenchel–Young loss $\lambda L_R(\mu \parallel \eta)$ below a tolerance $\epsilon > 0$, where $\eta := -\Phi^*h/\lambda$ is the candidate certificate associated to the dual iterate h . The classical Brøndsted–Rockafellar theorem produces a nearby exact subgradient pair, and Carlier [2023] makes this *constructive* by realizing the proxy as a proximal step. This dual-feasible proxy lifts to a primal certificate $\bar{\mu}$ with an *exact* source condition $\bar{\eta} \in \partial R(\bar{\mu})$, unconditionally for a norm regularizer (Proposition 3). Since nothing in the construction is tied to the quadratic geometry, we finally replace Carlier’s Euclidean prox by a mirror proximal step with kernel F^* tilted by the prediction $\Phi\mu$: it obeys an exact duality-gap identity (Proposition 5) and, in the quadratic case, improves the gap by exactly $\lambda L_R(\mu \parallel \eta)$.

Algorithmic implications. Mirror Alignment also relates to several heuristics used in deep-learning optimization. The continuous-time trajectories of the Lion-K framework [Chen et al., 2024] can be read as gradient flows of dual certificates of a continuous regularizer. The Muon optimizer [Jordan et al., 2024] performs matrix-sign (polar) orthogonalization of gradient updates onto the Stiefel manifold, which corresponds to the Mirror Alignment condition of a nuclear-norm penalty. Note that this polar step admits efficient implementations via the Polar Express algorithm [Amsel et al., 2026]. Section 5.2 demonstrates this experimentally: when Lion-K’s sign step and Muon’s matrix-sign step are used as the proximal operator of their regularizer (ℓ_1 and the nuclear norm), the resulting iterations coincide with ISTA (iterative shrinkage-thresholding) and SVT (singular-value thresholding), and hence solve the $F + \lambda R$ program. The Fenchel–Young duality-gap decomposition $\Delta = L_F + \lambda L_R$ (evaluated at a rescaled, dual-feasible certificate) certifies these iterates as shown in Figure 1 where the ill-conditioning of the problem makes convergence gradual: for both optimizers Δ_t decays steadily to zero and stays above the optimality gap, certifying (without any oracle access to the optimum) that they solve the $F + \lambda R$ program.

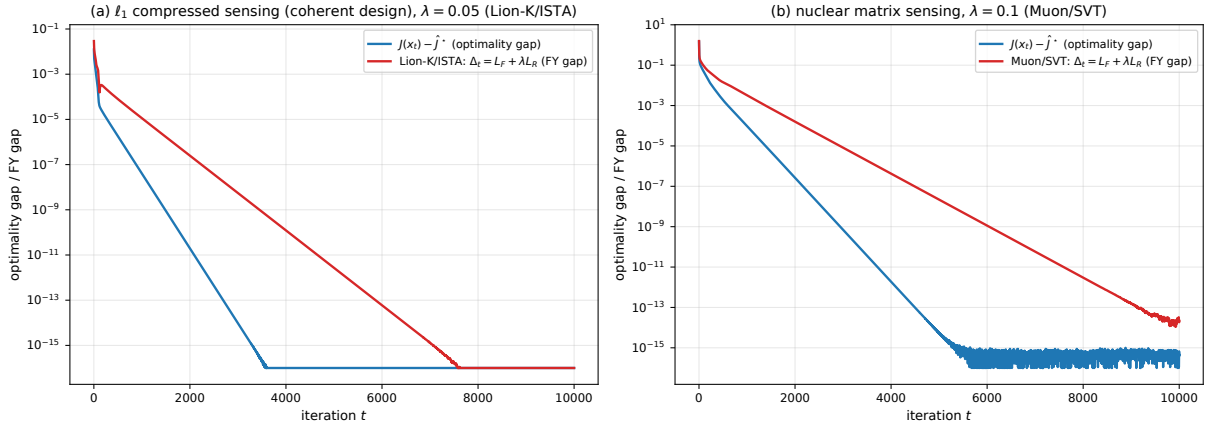


Figure 1: Fenchel–Young gap at a rescaled Mirror-Alignment dual, along the trajectories of two proximal $F + \lambda R$ solvers on ill-conditioned instances (double precision; first 10 000 iterations). **(a)** Coherent ℓ_1 compressed sensing (Lion-K / ISTA: the sign step used as the ℓ_1 proximal operator; AR(1)-correlated design with $\rho = 0.9$), penalty $\lambda = 0.05$. **(b)** Nuclear-norm matrix sensing from $p = 120$ Gaussian measurements (Muon / SVT: the matsign step used as the nuclear-norm proximal operator), penalty $\lambda = 0.1$. In each panel, red: the oracle-free Fenchel–Young gap Δ_t , which upper-bounds the (oracle) primal optimality gap $J(\mu_t) - \hat{J}^*$ shown in blue, with $\hat{J}^*$ the minimal objective over an extended reference run.

Companion paper. Building on the Fenchel–Young error bounds of this paper, a companion work [Aubin-Frankowski and De Castro, 2026] introduces a functional noise model and a non-degenerate source condition (NDSC; Azaïs et al., 2015, Duval and Peyré, 2015, Poon et al., 2023) under which the GBL solution is unique, exactly k -sparse with the correct signs, and converges at a strict first-order spatial rate in $(\lambda, \|w\|_{\mathcal{H}})$; it verifies the NDSC through a Fisher-reweighted Local Positive Curvature pivot, obtains sign-free support localization from any non-degenerate certificate via a Kernel Switch principle, and instantiates the theory on the well-specified population logistic GBL with the sinc-4 pivot, attaining the parametric statistical rate up to a logarithmic factor. The present paper supplies the gap decomposition and the error bounds on which that analysis rests.

1.2 Related work

The mathematical foundations of exact recovery over spaces of measures were formalized by Bredies and Pikkarainen [2013], who established the well-posedness of Tikhonov regularization with total-variation penalties on the space of finite Radon measures. The first quantitative error bounds for variational regularization with a *non-quadratic* convex penalty R are due to Burger and Osher [2004], who proved that under the source condition $\Phi^* \tilde{h} \in \partial R(\mu^*)$ the generalized Bregman divergence $D_R(\hat{\mu}_\lambda, \mu^*)$ between

the Tikhonov minimizer and the R -minimizing solution decays at rate $\mathcal{O}(\lambda + \delta)$ for $\lambda \asymp \delta$, with the explicit geometric specialization $D_R \equiv |\cdot|_{BV}$ on $BV(\Omega)$. Our framework recovers their identity as the regularizer-side component $\lambda L_R(\mu \parallel \eta^*)$ of the duality gap (Proposition 1) and generalizes the quadratic fidelity to arbitrary strictly convex F via Fenchel–Young losses. Continuous greedy algorithms (generalized conditional gradient and its sliding variants Boyd et al., 2017, Bredies and Pikkarainen, 2013, Denoyelle et al., 2019) give $\mathcal{O}(1/k)$ duality-gap rates for sparse deconvolution, but their analysis is tied almost exclusively to quadratic data fidelities.

Building on these conditions, the duality gap admits an exact decomposition into Fenchel–Young losses (see Blondel et al. [2020] for a review of the latter). The geometric utility of these losses was highlighted by Carlier [2023], who showed that a sharpened Fenchel–Young inequality controls the distance to the graph of the gradient mapping, ensuring measurement-space stability even when the primal is not unique; the same Brøndsted–Rockafellar bound (made *constructive* in the same paper) is what we use for early-stopping certification. Andrade et al. [2025] extended these sharpened losses to inverse entropic unbalanced optimal transport and to inverse Jordan–Kinderlehrer–Otto learning of population dynamics from independent snapshot samples, obtaining high-probability parameter recovery bounds. Other approximations of duality gaps exist, see Walwil and Fercoq [2025] for a recent review, here we stick to the analysis for the true gap, tying it to Bregman divergences and Fenchel–Young losses.

Notation. For a set, $\text{int}(\cdot)$ represents its topological interior. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed domain with $\mathcal{X} = \overline{\text{int} \mathcal{X}}$, possibly unbounded. We denote by $\mathcal{M}(\mathcal{X})$ the space of finite signed Radon measures on $\mathcal{X}$, equipped with the total variation norm $\|\cdot\|_{\text{TV}}$. Its pre-dual is the space $\mathcal{C}_0(\mathcal{X})$ of continuous functions vanishing at infinity, endowed with the supremum norm $\|\cdot\|_\infty$; for compact $\mathcal{X}$, $\mathcal{C}_0(\mathcal{X}) = \mathcal{C}(\mathcal{X})$. The duality pairing between $f \in \mathcal{C}_0(\mathcal{X})$ and $\mu \in \mathcal{M}(\mathcal{X})$ is denoted $\langle f, \mu \rangle_{\mathcal{C}, \mathcal{M}} = \int_{\mathcal{X}} f \, d\mu$; for a general pre-dual pair, $\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}$ denotes the pairing between $\mathcal{B}$ and $\mathcal{B}^*$. We let $\mathcal{H}$ denote a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. The forward operator is a linear map $\Phi : \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{H}$, continuous from the weak-* topology of $\mathcal{M}(\mathcal{X})$ to the weak topology of $\mathcal{H}$ (hence norm-bounded), with its pre-adjoint denoted by $\Phi^* : \mathcal{H} \rightarrow \mathcal{C}_0(\mathcal{X})$. For a convex function F , F^* denotes its Legendre–Fenchel conjugate, $\text{dom}(F)$ its domain and ∂F its subdifferential. For a closed convex set C , ι_C denotes its indicator function (zero on C and $+\infty$ outside) and, for $x \in C$, $N_C(x) := \partial \iota_C(x)$ its outward normal cone, that is the set of v in the dual of the ambient space with $\langle v, y - x \rangle \leq 0$ for every $y \in C$; for $C \subset \mathcal{H}$ the pairing is the inner product of $\mathcal{H}$. Furthermore, we denote the set of non-negative real numbers by $\mathbb{R}_+$. Given a discrete measure $\mu = \sum_{i=1}^k a_i \delta_{x_i}$, we denote its support by $S = \{x_1, \dots, x_k\}$ and the signs of its amplitudes by $s_i = \text{sign}(a_i)$. For $z \in \mathbb{R}^n$ and $\delta > 0$, $B_\infty^\delta(z) := \{w \in \mathbb{R}^n : \|w - z\|_\infty \leq \delta\}$ denotes the closed ℓ_∞ -ball of radius δ about z , and $B_\infty^\delta := B_\infty^\delta(0)$.

2 Fenchel duality, gap decomposition, and error bounds

This section starts with the GBL example, stating the source condition and Mirror Alignment informally, and then develops the abstract framework (duality-gap decomposition and Fenchel–Young error bounds) that makes them precise; the Brøndsted–Rockafellar algorithmic-stability theory, its primal-certificate lift, and a mirror-prox refinement follow in Section 3.

2.1 Setting and assumptions

To foster intuition, we begin with the GBL specialization (1.2, GBL), where the primal is a signed measure $\mu \in \mathcal{M}(\mathcal{X})$ and the regularizer is the total variation norm,

$$\inf_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ J_{\text{GBL}}(\mu) := F(\Phi\mu) + \lambda \|\mu\|_{\text{TV}} \right\}, \quad (2.1a)$$

and $F : \mathcal{H} \rightarrow \mathbb{R}$ is a strictly convex, continuously differentiable data-fidelity term and $\lambda > 0$ is the regularization parameter. Varying F encompasses several standard problems: *continuous sparse regression* (least squares; $F(h) = \frac{1}{2} \|h - y\|_{\mathcal{H}}^2$, the classical BLASSO of Bredies and Pikkarainen, 2013, De Castro and Gamboa, 2012); *classification BLASSO* (logistic or squared-hinge F); *counting/intensity estimation* (Poisson or KL); and *robust continuous sparse regression* (Huber). The full catalogue is treated in

Section 4. The Fenchel pre-dual of (2.1a) maximizes over an auxiliary variable $h \in \mathcal{H}$:

$$\sup_{h \in \mathcal{H}} \left\{ -F^*(h) \text{ subject to } \|\Phi^*h\|_\infty \leq \lambda \right\}. \quad (2.1b)$$

The constraint $\|\Phi^*h\|_\infty \leq \lambda$ enforces that the continuous function Φ^*h stays bounded uniformly on $\mathcal{X}$, preventing dual residuals from exceeding the sparsity threshold. Strong duality between (2.1a) and (2.1b) is established below as a consequence of the general F+R duality results.

The first-order condition for (2.1a) reads $0 \in \partial(F \circ \Phi + \lambda R)(\mu^*)$. Since the sum of subdifferentials is always contained in the subdifferential of the sum, it is enough to find μ^* such that $0 \in \partial(F \circ \Phi)(\mu^*) + \partial(\lambda R)(\mu^*)$; conversely, every primal solution admits such a dual variable (Theorem 2). As F is differentiable, $\partial(F \circ \Phi)(\mu^*) = \{\Phi^* \nabla F(\Phi \mu^*)\}$, so, introducing (i) $h^* = \nabla F(\Phi \mu^*)$ and (ii) $\eta^* \in \partial R(\mu^*)$, the problem becomes equivalent to having (i) and (ii) satisfied together with (iii) $\eta^* = -\Phi^*h^*/\lambda$. The combination of (ii) and (iii) asks the certificate $-\Phi^*h^*/\lambda$ to lie in $\partial\|\mu^*\|_{\text{TV}}$, and is known as the *source condition*; condition (i) is what we call *Mirror Alignment*. Both (i) and (ii) are equality cases of a Fenchel–Young inequality, that is, points of vanishing Fenchel–Young gap, which will be key for our reasoning below.

Source condition. The candidate certificate $\eta^* = -\Phi^*h^*/\lambda$ belongs to $\mathcal{C}_0(\mathcal{X})$. Concretely, for a k -sparse measure $\mu^* = \sum_{i=1}^k a_i^* \delta_{x_i^*}$ with signs $s_i = \text{sign}(a_i^*)$, the source condition asks that the continuous function η^* interpolates the signs on the support and is bounded by one everywhere:

$$\eta^*(x_i^*) = s_i \quad \text{for } i = 1, \dots, k, \quad \|\eta^*\|_\infty \leq 1.$$

Asking in addition the strict inequality $|\eta^*(x)| < 1$ off the support is the stronger *non-degenerate* source condition studied in the companion paper [Aubin-Frankowski and De Castro, 2026] and well known in the literature [Azaïs et al., 2015, Duval and Peyré, 2015, Poon et al., 2023].

Mirror Alignment. Condition (i) reads backwards as $\Phi \mu^* = \nabla F^*(h^*)$. We call this *Mirror Alignment*, because it generalizes the mirror-descent primal/dual identity beyond the squared loss.

Framework and notation. We work over a generic Banach-dual primal space $\mathcal{B}^*$ and a separable Hilbert measurement space $\mathcal{H}$. The framework rests on a triple of spaces and a pair of functions.

Definition 1 (Set $\mathcal{G}_0$). We define $\mathcal{G}_0$ as the set of triples $(\mathcal{H}, \mathcal{B}, \Phi)$ where $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a separable Hilbert space; $\mathcal{B}$ is a Banach space with dual space $\mathcal{B}^*$; and $\Phi : \mathcal{B}^* \rightarrow \mathcal{H}$ is a linear operator that is weak-*to-weak continuous, i.e. continuous from $(\mathcal{B}^*, \sigma(\mathcal{B}^*, \mathcal{B}))$ to $\mathcal{H}$ with its weak topology. The operator Φ is referred to as the forward operator.

By Lemma 8, the pre-adjoint $\Phi^* : \mathcal{H} \rightarrow \mathcal{B}$ maps into the pre-dual space $\mathcal{B}$, itself identified with a subspace of the bidual $\mathcal{B}^{**}$.

Remark 2 (The signed measures example). Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed set, possibly unbounded, and let $\mathcal{B} = (\mathcal{C}_0(\mathcal{X}), \|\cdot\|_\infty)$ be the space of continuous functions vanishing at infinity on $\mathcal{X}$ and $\mathcal{B}^* = (\mathcal{M}(\mathcal{X}), \|\cdot\|_{\text{TV}})$ its dual space, identified with the space of finite signed Radon measures on $\mathcal{X}$ by the Riesz–Markov theorem [Rudin, 1987, Theorem 6.19]; for compact $\mathcal{X}$ one has $\mathcal{C}_0(\mathcal{X}) = \mathcal{C}(\mathcal{X})$ and recovers the classical setting. By Lemma 9, for any linear weak-*to-weak continuous operator Φ , one has

$$\Phi \mu = \int_{\mathcal{X}} \varphi_t \, d\mu(t), \quad \mu \in \mathcal{M}(\mathcal{X}),$$

where $\varphi_t := \Phi \delta_t$ with δ_t the Dirac measure at $t \in \mathcal{X}$; and the dual operator $\Phi^* : \mathcal{H} \rightarrow \mathcal{C}_0(\mathcal{X})$ reads

$$\Phi^* : h \in \mathcal{H} \mapsto (t \in \mathcal{X} \mapsto \langle \varphi_t, h \rangle_{\mathcal{H}}) \in \mathcal{C}_0(\mathcal{X}),$$

where we identified the pre-dual space $\mathcal{C}_0(\mathcal{X})$ with a subspace of the dual $\mathcal{M}(\mathcal{X})^*$. For unbounded $\mathcal{X}$, membership of Φ^*h in $\mathcal{C}_0(\mathcal{X})$ is a real restriction on the forward operator: since $\delta_t \rightharpoonup^* 0$ as $|t| \rightarrow \infty$, weak-*to-weak continuity forces the features to vanish weakly, $\varphi_t \rightharpoonup 0$ in $\mathcal{H}$. Observables that do not vanish at infinity are excluded, such as the total mass $\mu \mapsto \mu(\mathcal{X}) = \langle \mathbf{1}, \mu \rangle$, admissible only for compact $\mathcal{X}$ since $\mathbf{1} \notin \mathcal{C}_0(\mathcal{X})$. Translation-invariant feature maps on $\mathcal{H} = L^2(\mathbb{R}^d)$ are however admissible, their translates converging weakly to zero. Among the examples of Section 4, only the total variation of the gradient (Section 4.2.4) requires a bounded domain.

Example 1 (Mixture of Gaussians). Take $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = L^2(\mathbb{R}^d)$ and let the feature φ_t be the Gaussian density on $\mathbb{R}^d$ centered at t , with covariance matrix $\sigma^2 \text{Id}$. The features are translates of one fixed function of $L^2(\mathbb{R}^d)$, so they converge weakly to zero as $|t| \rightarrow \infty$ and the triple $(\mathcal{H}, \mathcal{B}, \Phi)$ belongs to $\mathcal{G}_0$ even though $\mathcal{X} = \mathbb{R}^d$ is unbounded. The prediction

$$\Phi\mu = \int_{\mathbb{R}^d} \varphi_t d\mu(t)$$

is a mixture of Gaussians with mixing measure μ , and the GBL estimates μ from the mixture. With the squared norm $F = \frac{1}{2} \|\cdot - y\|_{\mathcal{H}}^2$ this is the BLASSO for mixture models of [De Castro et al. \[2021\]](#), where the data y is built from the empirical measure of a sample of the mixture and $\mathcal{H}$ is taken to be a reproducing kernel Hilbert space, so that this empirical measure can be compared with the prediction. The present paper covers the other data-fidelity terms of [Table 1](#) as well, in particular the logistic loss of [Section B.2](#), for which the same duality gap, certificates and error bounds hold.

We also need to consider the following functions.

Definition 3 (Set $\mathcal{F}_0$). Given $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, we define $\mathcal{F}_0$ as the set of pairs (F, R) where $F : \mathcal{H} \rightarrow \mathbb{R}$ is a continuously differentiable strictly convex function such that there exist $a > 0$ and $b \in \mathbb{R}$ with

$$\forall h \in \mathcal{H}, \quad F(h) \geq a\|h\|_{\mathcal{H}} + b; \quad (2.2)$$

$R : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper convex and weak-* lower semi-continuous function ($\sigma(\mathcal{B}^*, \mathcal{B})$ -l.s.c., equivalently $R = R^{**}$ with the conjugate R^* taken on the pre-dual $\mathcal{B}$), referred to as the regularizer. The function F is referred to as the data-fidelity term. We also write $F \in \mathcal{F}_0$ when F alone satisfies the conditions above on the data-fidelity side.

The lower bound (2.2) is a weak-coercivity condition: it makes F bounded below, so that the objective is bounded below and, together with the coercivity supplied by the regularizer, the $F + \lambda R$ program admits a minimizer. The norm-type regularizers in [Section 4](#) (total variation, group/joint TV, nuclear norm) are weak-* lower semi-continuous, as required in [Definition 3](#), being suprema of weak-* continuous linear functionals (e.g. $\|\mu\|_{\text{TV}} = \sup_{\|f\|_{\infty} \leq 1} \langle f, \mu \rangle$); the constrained regularizers (analysis sparsity, TV of the gradient) are weak-* l.s.c. provided the underlying constraint set is weak-* closed, see the standing hypothesis in [Section 4.2.3](#).

Now, consider the following convex program on the dual space $\mathcal{B}^*$

$$\inf_{\mu \in \mathcal{B}^*} \left\{ J_{\mathcal{B}^*}(\mu) := F(\Phi\mu) + \lambda R(\mu) \right\}, \quad (2.3)$$

where $J_{\mathcal{B}^*} : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ is the objective function.

Legendre–Fenchel duality. We define the *Legendre–Fenchel conjugates* of F and R as follows: for all $h \in \mathcal{H}$ and all $f \in \mathcal{B}$,

$$F^*(h) := \sup_{h' \in \mathcal{H}} \{ \langle h, h' \rangle_{\mathcal{H}} - F(h') \} \in \mathbb{R} \cup \{+\infty\} \quad (2.4a)$$

$$R^*(f) := \sup_{\mu \in \mathcal{B}^*} \{ \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R(\mu) \} \in \mathbb{R} \cup \{+\infty\} \quad (2.4b)$$

where R^* is only given on the pre-dual space $\mathcal{B}$, identified with a weak-* dense subset of the bidual $\mathcal{B}^{**}$.

Fenchel–Young losses. We also consider the *Fenchel–Young losses* [[Blondel et al., 2020](#)], defined as follows: for all $h, h' \in \mathcal{H}$ and all $(f, \mu) \in \mathcal{B} \times \mathcal{B}^*$,

$$L_F(h' \parallel h) := F(h') + F^*(h) - \langle h, h' \rangle_{\mathcal{H}} \in \mathbb{R}_+ \cup \{+\infty\}, \quad (2.5a)$$

$$L_R(\mu \parallel f) := R(\mu) + R^*(f) - \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} \in \mathbb{R}_+ \cup \{+\infty\}. \quad (2.5b)$$

Fenchel–Young losses can be understood as a generalization with primal–dual variables of the more common concept of Bregman divergence, written instead on primal–primal variables. We now recall this second notion and state the identity that links the two.

Bregman divergences. Given a convex Gâteaux-differentiable function $G : \mathcal{H} \rightarrow \mathbb{R}$, its Bregman divergence is, for all $x, y \in \mathcal{H}$,

$$G(x \mid y) := G(x) - G(y) - \langle \nabla G(y), x - y \rangle_{\mathcal{H}} \in \mathbb{R}_+.$$

For F and its convex conjugate F^* , we use for all $u, v \in \mathcal{H}$ and all $h \in \mathcal{H}$, $h' \in \text{int}(\text{dom}(F^*))$,

$$F(u \mid v) := F(u) - F(v) - \langle \nabla F(v), u - v \rangle_{\mathcal{H}} \in \mathbb{R}_+, \quad (2.6a)$$

$$F^*(h \mid h') := F^*(h) - F^*(h') - \langle \nabla F^*(h'), h - h' \rangle_{\mathcal{H}} \in \mathbb{R}_+ \cup \{+\infty\}. \quad (2.6b)$$

The function F is assumed differentiable and strictly convex so that F^* is also differentiable on $\text{int}(\text{dom}(F^*))$. Indeed, strict convexity of F ensures that the supremum defining $F^*(h)$ admits at most one maximizer. On $\text{int}(\text{dom}(F^*))$, where subgradients are guaranteed to exist, this uniqueness implies that F^* is Gâteaux differentiable; see Zălinescu [2002] for the infinite-dimensional Hilbert setting. The two notions then agree, for all $u \in \mathcal{H}$ and all $h \in \text{int}(\text{dom}(F^*))$,

$$F(u \mid \nabla F^*(h)) = F^*(h \mid \nabla F(u)) = L_F(u \parallel h), \quad (2.6c)$$

which is Lemma 11. Fenchel–Young losses have the advantage over Bregman divergences of not requiring differentiability of the functional.

2.2 Results on duality and optimality gaps

The *dual program* of (2.3) (strictly speaking its Fenchel *pre-dual*, since $\mathcal{B}^*$ is itself a dual space) reads

$$\sup_{h \in \mathcal{H}} \left\{ J_{\mathcal{H}}(h) := -F^*(h) - \lambda R^* \left(-\frac{\Phi^* h}{\lambda} \right) \right\}, \quad (2.7)$$

where $J_{\mathcal{H}} : \mathcal{H} \rightarrow \mathbb{R} \cup \{-\infty\}$ is the objective function.

Duality-gap decomposition. Theorem 1 below expresses the duality gap in terms of Fenchel–Young losses.

Theorem 1 (Fenchel–Young losses and duality gap). *Let $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, let $F : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ be proper, l.s.c. convex, let $R : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ be proper convex weak-* l.s.c., and let $\lambda > 0$. Fix $\mu \in \text{dom}(R) \subset \mathcal{B}^*$ with $\Phi\mu \in \text{dom}(F)$, and $h \in \text{dom}(F^*) \subset \mathcal{H}$. Then the duality gap $\Delta(\mu, h)$ satisfies*

$$\Delta(\mu, h) := J_{\mathcal{B}^*}(\mu) - J_{\mathcal{H}}(h) = L_F(\Phi\mu \parallel h) + \lambda L_R(\mu \parallel \eta) \in \mathbb{R}_+ \cup \{+\infty\}. \quad (2.8)$$

where $\eta := -\Phi^* h / \lambda$.

When $(F, R) \in \mathcal{F}_0$ the requirement $\Phi\mu \in \text{dom}(F)$ is automatic, since $\text{dom}(F) = \mathcal{H}$. Note that the gap is finite (and the identity holds in $\mathbb{R}_+$) precisely when $\eta \in \text{dom}(R^*)$, equivalently $h \in \text{dom}(J_{\mathcal{H}})$; otherwise $L_R(\mu \parallel \eta) = +\infty$, $J_{\mathcal{H}}(h) = -\infty$, and both sides equal $+\infty$.

Proof. First, we verify that the functional $J_{\mathcal{H}}$ defined in (2.7) corresponds to the Fenchel pre-dual of the primal problem (2.3). We write the Fenchel–Rockafellar dual (see Zălinescu, 2002, Section 2.8) of the objective (2.3) rewritten as $\inf_{\mu} \{F(\Phi\mu) + g_{\lambda}(\mu)\}$, where $g_{\lambda}(\mu) := \lambda R(\mu)$:

$$\sup_{h \in \mathcal{H}} \left\{ -F^*(h) - g_{\lambda}^*(-\Phi^* h) \right\}.$$

We compute the conjugate of the scaled regularizer $g_{\lambda}(\mu) = \lambda R(\mu)$, giving $g_{\lambda}^*(v) = \lambda R^*(v/\lambda)$; substituting into the dual formulation, we match exactly the definition of $J_{\mathcal{H}}(h)$ in (2.7). Thus, $J_{\mathcal{H}}$ is the canonical (pre-)dual.¹

If $\eta = -\Phi^* h / \lambda \notin \text{dom}(R^*)$, then $R^*(\eta) = +\infty$, hence $L_R(\mu \parallel \eta) = +\infty$ and $J_{\mathcal{H}}(h) = -\infty$, so $\Delta(\mu, h) = +\infty$ matches the right-hand side and the identity holds in $\mathbb{R}_+ \cup \{+\infty\}$. Assume henceforth $\eta \in \text{dom}(R^*)$,

¹One can start from (2.7) and, by the same algebraic manipulations on the Legendre–Fenchel transform together with $F = F^{**}$ and $R = R^{**}$ (weak-* l.s.c.), deduce that (2.3) is the dual problem to (2.7). For the sake of readability, we present its algebraic proof the other way around using implicitly that $\mathcal{B} \hookrightarrow \mathcal{B}^{**}$.

so all terms below are finite. We add and subtract the same terms to recognize the Fenchel–Young losses and we use Lemma 8:

$$\begin{aligned} J_{\mathcal{B}^*}(\mu) &= F(\Phi\mu) + \lambda R(\mu) \\ &= F(\Phi\mu) + F^*(h) - \langle h, \Phi\mu \rangle_{\mathcal{H}} + \lambda(R(\mu) + R^*(\eta) - \langle \eta, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}) - F^*(h) - \lambda R^*(\eta) \\ &= L_F(\Phi\mu \parallel h) + \lambda L_R(\mu \parallel \eta) + J_{\mathcal{H}}(h), \end{aligned}$$

as claimed, where $\eta = -\Phi^*h/\lambda$. $\square$

Remark 4 (Scope of the gap identity). *The gap identity of Theorem 1 uses none of the differentiability, strict convexity or coercivity of F : the proof only adds and subtracts conjugate terms. Section 4 uses this generality for the losses outside $\mathcal{F}_0$ (logistic, Huber, squared hinge, linear-link KL).*

The identity (2.8) has direct consequences. By weak duality, it provides an upper bound on the optimality gap of $J_{\mathcal{B}^*}$: for all $\mu \in \text{dom}(R)$ with $\Phi\mu \in \text{dom}(F)$ and all $h \in \text{dom}(F^*)$,

$$0 \leq J_{\mathcal{B}^*}(\mu) - \inf_{\nu \in \mathcal{B}^*} J_{\mathcal{B}^*}(\nu) \leq \inf_{h' \in \mathcal{H}} \Delta(\mu, h') \leq \Delta(\mu, h) \stackrel{(2.8)}{=} L_F(\Phi\mu \parallel h) + \lambda L_R(\mu \parallel \eta). \quad (2.9)$$

Since the duality gap (2.8) is the sum of two non-negative Fenchel–Young losses, driving it to zero requires both terms to vanish. By a standard result recalled in Lemma 10, the condition $L_R(\mu \parallel \eta) = 0$ is equivalent to the subdifferential inclusion $\eta \in \partial R(\mu)$. In the context of inverse problems, this inclusion governs the structural properties of the solution, such as the sparsity promoted by the total variation norm. This motivates the following definition of the source condition.

Definition 5 (Dual certificate/source condition). *Let $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, $R : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ a proper convex weak-* l.s.c. function, $\lambda > 0$, and $\mu \in \mathcal{B}^*$. We say that $\eta \in \mathcal{B}$ is a dual certificate of μ for $(\mathcal{H}, \lambda, R, \Phi)$ if it satisfies the source condition at point μ : there exists $h \in \mathcal{H}$ such that*

$$\eta = -\frac{\Phi^*h}{\lambda} \in \partial R(\mu), \quad (2.10)$$

where ∂R denotes the subdifferential of R , understood through the pairing $(\mathcal{B}, \mathcal{B}^*)$: we only consider subgradients lying in $\mathcal{B}$.

We can state the following theorem.

Theorem 2 (First-order optimality and Mirror Alignment). *Let $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, $(F, R) \in \mathcal{F}_0$ and $\lambda > 0$. Let $\mu \in \mathcal{B}^*$ and let $h \in \mathcal{H}$ be such that the source condition (2.10) holds. Then*

$$0 \leq J_{\mathcal{B}^*}(\mu) - \inf_{\nu \in \mathcal{B}^*} J_{\mathcal{B}^*}(\nu) \leq L_F(\Phi\mu \parallel h),$$

and, when $h \in \text{int}(\text{dom } F^*)$, $L_F(\Phi\mu \parallel h) = F(\Phi\mu \mid \nabla F^*(h)) = F^*(h \mid \nabla F(\Phi\mu))$. If additionally $h \in \mathcal{H}$ satisfies one of these equivalent conditions

$$h = \nabla F(\Phi\mu) \quad (2.11a)$$

$$\Phi\mu = \nabla F^*(h) \quad (2.11b)$$

then μ is a solution to (2.3) and h a solution to (2.7). In (2.11), $\nabla F^*(h)$ denotes the unique subgradient of F^* at h : by strict convexity of F , $\partial F^*(h)$ contains at most one element, and it is nonempty whenever $h = \nabla F(\Phi\mu)$.

The converse is true: If μ is a solution to (2.3), then there exists a solution $h \in \mathcal{H}$ to (2.7) such that (2.10) holds and (2.11) holds (Mirror Alignment).

Proof. Set $\eta := -\Phi^*h/\lambda$. By the source condition $\eta \in \partial R(\mu)$, so $L_R(\mu \parallel \eta) = 0$ (Lemma 10), and the gap decomposition of Theorem 1 gives $\Delta(\mu, h) = L_F(\Phi\mu \parallel h)$. By weak duality, $0 \leq J_{\mathcal{B}^*}(\mu) - \inf_{\nu \in \mathcal{B}^*} J_{\mathcal{B}^*}(\nu) \leq \Delta(\mu, h) = L_F(\Phi\mu \parallel h)$, which is the first inequality. Moreover, by the zero-loss equivalence of Fenchel–Young losses (Lemma 10), $L_F(\Phi\mu \parallel h) = 0$ iff $h = \nabla F(\Phi\mu)$ (Mirror Alignment) and $L_R(\mu \parallel \eta) = 0$ iff $\eta \in \partial R(\mu)$ (the source condition); the two conditions together are therefore equivalent to $\Delta(\mu, h) = 0$, i.e. to optimality of (μ, h) .

Conversely, suppose μ is a solution to (2.3). Since $(F, R) \in \mathcal{F}_0$, the fidelity F is finite and continuous on $\mathcal{H}$ and $\text{dom } R \neq \emptyset$, so F is continuous at $\Phi\mu_0$ for some $\mu_0 \in \text{dom } R$ and the qualification of the Fenchel–Rockafellar duality theorem [Zălinescu, 2002, Corollary 2.8.5] holds; it yields strong duality together with attainment of the dual supremum in (2.7) (the weak-coercivity bound (2.2) is not used here). Hence there is $h \in \mathcal{H}$ with $\Delta(\mu, h) = 0$.

By the gap identity of Theorem 1 and the non-negativity of the Fenchel–Young losses, $0 = \Delta(\mu, h) = L_F(\Phi\mu \parallel h) + \lambda L_R(\mu \parallel \eta)$ with $\eta = -\Phi^*h/\lambda$ forces $L_F(\Phi\mu \parallel h) = 0$ and $L_R(\mu \parallel \eta) = 0$; the zero-loss equivalences used above then give Mirror Alignment $h = \nabla F(\Phi\mu)$ and the source condition $\eta \in \partial R(\mu)$, so h solves (2.7). $\square$

Remark 6 (Mirror Alignment). *The optimality conditions of Theorem 2, $L_R = 0$ (the source condition (2.10)) and $L_F = 0$ (Mirror Alignment (2.11)), recover the classical inverse-problem conditions [Bredies and Pikkarainen, 2013]: the source condition is the subgradient inclusion $-\Phi^*h/\lambda \in \partial R(\mu)$. Our formulation highlights Mirror Alignment $h = \nabla F(\Phi\mu)$ (equivalently $\Phi\mu = \nabla F^*(h)$): whereas first-order conditions present the primal–dual relation algebraically (linearly in the quadratic case), Mirror Alignment reads it as the Legendre–Fenchel conjugacy induced by F . This underlies the Fenchel–Young error bounds of Theorem 3 below, since L_F coincides with a Bregman divergence of F under Mirror Alignment (Lemma 11).*

The source condition (2.10) is a sufficient condition for the Fenchel–Young loss of R to coincide with its generalized Bregman divergence at the certified point,

$$R_{\eta^*}(\mu \mid \mu^*) := R(\mu) - R(\mu^*) - \langle \eta^*, \mu - \mu^* \rangle_{\mathcal{B}, \mathcal{B}^*}$$

for $\eta^* \in \partial R(\mu^*)$, written D_R in Burger and Osher [2004]. The converse direction (Brøndsted–Rockafellar, Theorem 4) shows that an ϵ -bounded L_R already yields a nearby exact certificate.

Proposition 1 (Regularizer Fenchel–Young loss equals the Bregman divergence under the source condition). *Let $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, $R : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ a proper convex weak-* l.s.c. function, $\lambda > 0$, $\mu^* \in \mathcal{B}^*$ and $\eta^* \in \mathcal{B}$ a dual certificate of μ^* for $(\mathcal{H}, \lambda, R, \Phi)$ (i.e. (2.10) holds at μ^*). Then*

$$\forall \mu \in \mathcal{B}^*, \quad L_R(\mu \parallel \eta^*) = R_{\eta^*}(\mu \mid \mu^*). \quad (2.12)$$

Proof. By the source condition, $\eta^* \in \partial R(\mu^*)$, so $L_R(\mu^* \parallel \eta^*) = 0$ (see Lemma 10), i.e. $R^*(\eta^*) = \langle \eta^*, \mu^* \rangle_{\mathcal{B}, \mathcal{B}^*} - R(\mu^*)$. Substituting into the Fenchel–Young loss (2.5b),

$$L_R(\mu \parallel \eta^*) = R(\mu) + R^*(\eta^*) - \langle \eta^*, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = R(\mu) - R(\mu^*) - \langle \eta^*, \mu - \mu^* \rangle_{\mathcal{B}, \mathcal{B}^*} = R_{\eta^*}(\mu \mid \mu^*),$$

the generalized Bregman divergence, which is (2.12). $\square$

2.3 Error bounds

We now turn the duality-gap identity into a priori error bounds. Fix $\mu^* \in \text{dom}(R)$ and consider *any* $h^* \in \text{dom}(F^*)$ satisfying the source condition (2.10) at μ^* , that is

$$\text{Certificate:} \quad \eta^* := -\frac{\Phi^*h^*}{\lambda} \in \partial R(\mu^*).$$

The message of this subsection is that a single scalar, given by

$$\text{Error upper bound:} \quad L_F(\Phi\mu^* \parallel h^*),$$

controls the recovery errors of any near minimizer μ (making an improvement over μ^*) in the sense that

$$\text{Set of near minimizers:} \quad J_{\mathcal{B}^*}(\mu) \leq J_{\mathcal{B}^*}(\mu^*).$$

Theorem 3 states an exact identity, the resulting bound, and its split into an estimation part and a prediction part.

Theorem 3 (Fenchel–Young error bounds). *Let $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, let $F : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ be proper, lower semi-continuous and convex, let $R : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ be proper convex and weak-* l.s.c., and $\lambda > 0$. Let $\mu^* \in \text{dom}(R)$ and $h^* \in \text{dom}(F^*)$ satisfy the source condition (2.10) at μ^* , with $J_{\mathcal{B}^*}(\mu^*) < +\infty$, and set $\eta^* := -\Phi^* h^* / \lambda$. Then, for all $\mu \in \text{dom}(R)$ with $\Phi\mu \in \text{dom}(F)$, the following holds.*

- One has

$$J_{\mathcal{B}^*}(\mu^*) - J_{\mathcal{B}^*}(\mu) = L_F(\Phi\mu^* \parallel h^*) - L_F(\Phi\mu \parallel h^*) - \lambda L_R(\mu \parallel \eta^*), \quad (2.13a)$$

where $L_R(\mu \parallel \eta^*)$ equals the Bregman divergence $R_{\eta^*}(\mu \mid \mu^*)$.

- If moreover $J_{\mathcal{B}^*}(\mu) \leq J_{\mathcal{B}^*}(\mu^*)$, then the Fenchel–Young error bounds hold:

$$0 \leq L_F(\Phi\mu \parallel h^*) + \lambda L_R(\mu \parallel \eta^*) \leq L_F(\Phi\mu^* \parallel h^*). \quad (2.13b)$$

- In particular for $J_{\mathcal{B}^*}(\mu) \leq J_{\mathcal{B}^*}(\mu^*)$, as $\max\{\lambda L_R(\mu \parallel \eta^*), L_F(\Phi\mu \parallel h^*)\} \leq L_F(\Phi\mu^* \parallel h^*)$, the two bounds below hold and are controlled simultaneously: the estimation bound (on μ)

$$0 \leq \lambda L_R(\mu \parallel \eta^*) \leq L_F(\Phi\mu^* \parallel h^*) \quad (2.13c)$$

and the prediction bound (on $\Phi\mu$)

$$0 \leq L_F(\Phi\mu \parallel h^*) \leq L_F(\Phi\mu^* \parallel h^*). \quad (2.13d)$$

- Finally, for any $\tau > 0$,

$$\frac{1}{2\tau} \text{dist}_{\mathcal{H} \times \mathcal{H}}((\Phi\mu^*, \tau h^*), \text{Gr}(\tau \partial F))^2 \leq \frac{1}{\tau} \|\Phi\mu^* - (\text{Id} + \tau \partial F)^{-1}(\Phi\mu^* + \tau h^*)\|_{\mathcal{H}}^2 \leq L_F(\Phi\mu^* \parallel h^*), \quad (2.13e)$$

where $\text{Gr}(\tau \partial F)$ denotes the graph in $\mathcal{H} \times \mathcal{H}$ of $\tau \partial F$, and $(\text{Id} + \tau \partial F)^{-1}$ is the single-valued resolvent of the maximal monotone ∂F ; for a smooth F , $\partial F = \{\nabla F\}$.

Note that (2.13e) is the sharpened Fenchel–Young inequality of [Carlier \[2023, Lemma 1.1\]](#) which bounds the scalar $L_F(\Phi\mu^* \parallel h^*)$ from below by a squared distance to the graph of the subdifferential ∂F (the gradient ∇F when F is smooth).

Proof. By the source condition, $L_R(\mu^* \parallel \eta^*) = 0$ (Lemma 10), so $R^*(\eta^*)$ is finite and $h^* \in \text{dom}(J_{\mathcal{H}})$. The gap identity of Theorem 1, applied at (μ^*, h^*) and at (μ, h^*) gives, the finite dual value $J_{\mathcal{H}}(h^*)$ cancelling,

$$J_{\mathcal{B}^*}(\mu^*) - J_{\mathcal{B}^*}(\mu) = \Delta(\mu^*, h^*) - \Delta(\mu, h^*) = L_F(\Phi\mu^* \parallel h^*) - L_F(\Phi\mu \parallel h^*) - \lambda L_R(\mu \parallel \eta^*),$$

which is (2.13a); the identification with the Bregman divergence stems from Proposition 1. If $J_{\mathcal{B}^*}(\mu) \leq J_{\mathcal{B}^*}(\mu^*)$, the left-hand side of (2.13a) is non-negative and the upper bound in (2.13b) follows; the lower bound holds because both Fenchel–Young losses are non-negative (Lemma 10).

For (2.13e), the subdifferential ∂F is maximal monotone (F proper, l.s.c. and convex, by Rockafellar’s theorem), so the resolvent $(\text{Id} + \tau \partial F)^{-1}$ is single-valued and defined on all of $\mathcal{H}$ (Minty); for a smooth F it is $(\text{Id} + \tau \nabla F)^{-1}$. [Carlier \[2023, Lemma 1.1\]](#) applied to F at $(\Phi\mu^*, h^*)$ gives $L_F(\Phi\mu^* \parallel h^*) \geq \tau^{-1} \|\Phi\mu^* - \bar{x}\|_{\mathcal{H}}^2$ with $\bar{x} := (\text{Id} + \tau \partial F)^{-1}(\Phi\mu^* + \tau h^*)$.

Set $\bar{p} := h^* + \tau^{-1}(\Phi\mu^* - \bar{x}) \in \partial F(\bar{x})$ (the resolvent identity). Then, one has $\tau \|h^* - \bar{p}\|_{\mathcal{H}} = \|\Phi\mu^* - \bar{x}\|_{\mathcal{H}}$ and $(\bar{x}, \tau \bar{p}) \in \text{Gr}(\tau \partial F)$, so the lower bound equals $\frac{1}{2\tau} (\|\Phi\mu^* - \bar{x}\|_{\mathcal{H}}^2 + \tau^2 \|h^* - \bar{p}\|_{\mathcal{H}}^2) \geq \frac{1}{2\tau} \text{dist}_{\mathcal{H} \times \mathcal{H}}((\Phi\mu^*, \tau h^*), \text{Gr}(\tau \partial F))^2$.

Dropping either of the two non-negative terms on the left of (2.13b) gives (2.13c) and (2.13d); both being non-negative, their maximum is at most their sum, hence at most $L_F(\Phi\mu^* \parallel h^*)$. $\square$

Interpretation as error bounds. Theorem 3 bounds the estimation error (measured by λL_R) and the prediction error (measured by L_F) by the single scalar $L_F(\Phi\mu^* \parallel h^*)$. Under the source condition this scalar is the duality gap at the certified pair, $L_F(\Phi\mu^* \parallel h^*) = \Delta(\mu^*, h^*)$. It vanishes exactly when Mirror Alignment holds at the certificate, $h^* \in \partial F(\Phi\mu^*)$ (Lemma 10), the single-valued $h^* = \nabla F(\Phi\mu^*)$ under $\mathcal{F}_0$; in that case both recovery errors are zero.

Remark 7 (Best certificate and bounds without the source condition). (i) The bounds above are tightest for the certificate

$$h^* \in \arg \min_{h \in \mathcal{H}} \left\{ L_F(\Phi \mu^* \parallel h); -\frac{\Phi^* h}{\lambda} \in \partial R(\mu^*) \right\}.$$

One can check that a minimizer exists provided $\Phi \mu^* \in \text{int dom } F$ (in particular whenever $(F, R) \in \mathcal{F}_0$, where $\text{dom } F = \mathcal{H}$) and it is unique whenever F is differentiable (in particular under $\mathcal{F}_0$) and one minimizer lies in $\text{int dom } F^*$.

(ii) Without the source condition, Theorem 1 still gives error bounds. Let $\mu, \mu^* \in \text{dom}(R)$ and $h \in \text{dom}(J_{\mathcal{H}})$, and set $\eta := -\Phi^* h / \lambda$. Subtracting the gap identity at (μ^*, h) and (μ, h) gives the identity $J_{\mathcal{B}^*}(\mu^*) - J_{\mathcal{B}^*}(\mu) = \Delta(\mu^*, h) - \Delta(\mu, h)$, so $J_{\mathcal{B}^*}(\mu) \leq J_{\mathcal{B}^*}(\mu^*)$ implies

$$0 \leq L_F(\Phi \mu \parallel h) + \lambda L_R(\mu \parallel \eta) \leq \Delta(\mu^*, h) = L_F(\Phi \mu^* \parallel h) + \lambda L_R(\mu^* \parallel \eta).$$

The bound $\Delta(\mu^*, h)$ now carries the extra regularizer term; it reduces to the single scalar of Theorem 3 exactly under the source condition.

3 Approximate source conditions and algorithmic stability

Section 2 characterized optimality through the source condition $\eta^* \in \partial R(\mu^*)$ holding *exactly*. Practical continuous optimization algorithms (conditional gradient methods, Forward–Backward splitting, or non-convex particle flows like Lion-K) are subject to discretization error, finite stopping times, and non-convex landscapes, so they typically do not satisfy this inclusion. Early stopping yields instead an ϵ -approximate source condition measured directly by the regularizer Fenchel–Young loss, as $\lambda L_R(\mu \parallel \eta) \leq \epsilon$, where ϵ depends (not exclusively) on the number of (conditional) gradient steps, the dimension and the initialization, as described by the available rates of convergence of the chosen algorithm. We then show that any such ϵ -approximate primal–dual pair can be upgraded to a *dual-feasible* proxy. The classical theorem by Brøndsted and Rockafellar [1965], stated on Banach spaces, guarantees the *existence* of a nearby exact subgradient pair based on an order argument via Zorn’s lemma and Bishop–Phelps lemma. Carlier [2023] makes the argument *constructive* on Hilbert spaces, realizing the proxy as a proximal step: it lies within Hilbert distance $\sqrt{\tau \epsilon}$ of the iterate for every free balancing parameter $\tau > 0$ (Theorem 4), and the Fenchel–Young tolerance becomes an early-stopping criterion. We state the result for an arbitrary proper convex weak-* lower semi-continuous ($\sigma(\mathcal{B}^*, \mathcal{B})$ -l.s.c.) regularizer R and specialize it to the GBL total-variation norm afterwards (Remark 9). Section 3.3 then refines the construction in the geometry of the data-fidelity term: the Euclidean prox is replaced by a mirror proximal step with kernel F^* , tilted by the prediction $\Phi \mu$, which recovers Carlier’s step when F is quadratic and ties the proxy to the Mirror Alignment of Section 2.

3.1 Hilbert pull-back of the regularizer

Carlier’s *constructive* Brøndsted–Rockafellar theorem [Carlier, 2023, Theorem 4.1] is stated for convex functions on a Hilbert space, whereas R acts on the Banach space $\mathcal{B}^*$. We hence pull R back to $\mathcal{H}$ through Φ^* so that Carlier’s result applies (Theorem 4).

Assumption 1 (Regularizer admissible for Brøndsted–Rockafellar). Assume that $(\mathcal{H}, \mathcal{B}, \Phi) \in \mathcal{G}_0$, the regularizer $R : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper, convex and weak-* lower semi-continuous, and $\lambda > 0$. Define the pulled-back regularizer

$$u : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}, \quad u(h) := \lambda R^*\left(-\frac{\Phi^* h}{\lambda}\right), \quad \text{dom } u = \left\{h \in \mathcal{H} : -\frac{\Phi^* h}{\lambda} \in \text{dom } R^*\right\}. \quad (3.1)$$

We assume $\text{dom } u \neq \emptyset$.

The pulled-back regularizer u in (3.1) is convex, l.s.c. and proper on $\mathcal{H}$; see Lemma 12 for the proof. The feasibility requirement $\text{dom } u \neq \emptyset$ holds whenever $0 \in \text{dom } R^*$, in particular for every norm regularizer, where $R^* = \iota_{B_{\mathcal{B}}}$ is the indicator of the unit ball $B_{\mathcal{B}}$ of the pre-dual and 0 lies inside it, so $0 \in \text{dom } u$.

Proposition 2 (Hilbert pull-back and gap inequality). *Under Assumption 1, for every $\mu \in \text{dom } R$ and every $h \in \text{dom } u$, writing $\eta := -\Phi^*h/\lambda$, one has $u^*(-\Phi\mu) \leq \lambda R(\mu)$ and hence*

$$L_u(h \parallel -\Phi\mu) = u(h) + u^*(-\Phi\mu) - \langle h, -\Phi\mu \rangle_{\mathcal{H}} \leq \lambda L_R(\mu \parallel \eta). \quad (3.2)$$

Equality holds iff the supremum over the subspace already reaches the full value, i.e. iff

$$\sup_{f \in \text{ran } \Phi^*} [\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R^*(f)] = R(\mu).$$

Note that a sufficient condition for equality in (3.2) is that μ admits a dual certificate in $\text{ran } \Phi^*$, that is some $f \in \text{ran } \Phi^*$ with $f \in \partial R(\mu)$ (Lemma 13(i)). For a norm regularizer $R = \|\cdot\|_{\mathcal{B}^*}$, in particular $R = \|\cdot\|_{\text{TV}}$, a certificate in the closure $\overline{\text{ran } \Phi^*}$ is already enough, by rescaling (Lemma 13(ii)). The converse of the range condition is then false: if $\text{ran } \Phi^*$ is dense in $\mathcal{B}$ (a non-closed $\text{ran } \Phi^*$ is the generic situation for a compact Φ), equality holds for every $\mu \in \text{dom } R$ while there may be no certificate (Lemma 13(iii)).

Proof. By definition,

$$u^*(-\Phi\mu) = \sup_{h' \in \mathcal{H}} \langle -\Phi\mu, h' \rangle_{\mathcal{H}} - \lambda R^*(-\Phi^*h'/\lambda).$$

Set $f = -\Phi^*h'/\lambda$; then

$$\langle -\Phi\mu, h' \rangle_{\mathcal{H}} = -\langle \mu, \Phi^*h' \rangle_{\mathcal{B}^*, \mathcal{B}} = \lambda \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*},$$

and as h' ranges over $\mathcal{H}$, f ranges over the subspace $\Phi^*(\mathcal{H})$. Therefore

$$u^*(-\Phi\mu) = \lambda \sup_{f \in \Phi^*(\mathcal{H})} [\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R^*(f)] \leq \lambda \sup_{f \in \mathcal{B}} [\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R^*(f)] = \lambda R^{**}(\mu) \stackrel{(*)}{=} \lambda R(\mu),$$

with equality iff the supremum over the subspace $\Phi^*(\mathcal{H})$ already equals the unconstrained one; here $(*)$ is the Fenchel–Moreau identity $R^{**} = R$ for the dual pair $(\mathcal{B}, \mathcal{B}^*)$, valid since R is proper, convex and weak-* l.s.c. (Assumption 1), with R^* taken on the pre-dual $\mathcal{B}$ so that no reflexivity of $\mathcal{B}$ is used (cf. Lemma 10). Finally, using $u(h) = \lambda R^*(\eta)$ and $\langle h, -\Phi\mu \rangle_{\mathcal{H}} = -\langle \Phi^*h, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = \lambda \langle \eta, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}$,

$$L_u(h \parallel -\Phi\mu) = \lambda R^*(\eta) + u^*(-\Phi\mu) - \lambda \langle \eta, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} \leq \lambda (R(\mu) + R^*(\eta) - \langle \eta, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}) = \lambda L_R(\mu \parallel \eta). \quad \square$$

Remark 8 (Why the gap is evaluated at $(h, -\Phi\mu)$). *The slack between $L_u(h \parallel -\Phi\mu)$ and $\lambda L_R(\mu \parallel \eta)$ measures the failure of topological certifiability, i.e. the gap between $u^*(-\Phi\mu)$ and $\lambda R(\mu)$; it is generically non-zero when $\Phi^*(\mathcal{H})$ is not sup-norm dense in $\mathcal{B}$. The inequality direction is exactly what we need: an upper bound on λL_R propagates to an upper bound on L_u .*

3.2 Algorithmic stability

In this section we show that any ϵ -approximate primal–dual pair (μ, h) can be upgraded to a nearby *dual-feasible* proxy $(\bar{\mu}, \bar{h})$ satisfying the exact subgradient inclusion $\bar{\eta} \in \partial R(\bar{\mu})$, with $\bar{\eta} = -\Phi^*\bar{h}/\lambda$. The construction proceeds in two steps: a Brøndsted–Rockafellar proximal step on the pulled-back regularizer u of (3.1), within Hilbert distance set by the Fenchel–Young tolerance ϵ (Theorem 4); and its lift to a primal certificate $\bar{\mu}$ (Proposition 3). Nothing in that step is specific to the quadratic geometry, and Section 3.3 builds the refinement adapted to F^* . For all $\tau > 0$ the proximal operator of τu is defined by

$$\text{prox}_{\tau u}(h) := \arg \min_{h' \in \mathcal{H}} \left\{ u(h') + \frac{1}{2\tau} \|h' - h\|_{\mathcal{H}}^2 \right\} = \arg \min_{h' \in \mathcal{H}} \left\{ \frac{1}{2\tau} \|h' - h\|_{\mathcal{H}}^2 + \lambda R^*\left(-\frac{\Phi^*h'}{\lambda}\right) \right\}.$$

Theorem 4 (Algorithmic stability via Brøndsted–Rockafellar, general regularization). *Let Assumption 1 hold and suppose an algorithm produces an approximate primal–dual pair $(\mu, h) \in \text{dom } R \times \text{dom } u$ whose regularizer Fenchel–Young loss obeys*

$$\lambda L_R(\mu \parallel \eta) \leq \epsilon, \quad \text{where } \eta = -\frac{\Phi^*h}{\lambda}. \quad (3.3a)$$

Then for every $\tau > 0$ there exist $\bar{h} \in \text{dom } u$ and $\bar{v} \in \partial u(\bar{h}) \subseteq \mathcal{H}$ given by

$$\bar{h} := \text{prox}_{\tau u}(h - \tau \Phi\mu), \quad \bar{v} := -\Phi\mu + \tau^{-1}(h - \bar{h}), \quad (3.3b)$$

and satisfying

$$\|\bar{h} - h\|_{\mathcal{H}} \leq \sqrt{\tau\epsilon}, \quad \|\bar{v} - (-\Phi\mu)\|_{\mathcal{H}} \leq \sqrt{\epsilon/\tau}. \quad (3.3c)$$

Note that the proxy is *dual-feasible*, $\bar{\eta} := -\Phi^* \bar{h} / \lambda \in \text{dom } R^*$, and $\bar{v} \in \partial u(\bar{h})$ is an *exact* subgradient inclusion for the pulled-back regularizer u of (3.1). Note also that the tolerance ϵ need not be small: one may take $\epsilon = \lambda L_R(\mu \parallel \eta)$, the running regularizer loss itself, so that the bounds (3.3c) hold at any iterate with no smallness assumption.

Proof. By Proposition 2 and (3.3a), $L_u(h \parallel -\Phi\mu) \leq \lambda L_R(\mu \parallel \eta) \leq \epsilon$. We apply Carlier's constructive Brøndsted–Rockafellar theorem [Carlier, 2023, Theorem 4.1] to the convex l.s.c. proper function u of (3.1) at $(x, p) = (h, -\Phi\mu)$: for every $\tau > 0$ the proximal point $\bar{h} = \text{prox}_{\tau u}(h - \tau\Phi\mu)$ and $\bar{v} = -\Phi\mu + \tau^{-1}(h - \bar{h})$ form a graph point $(\bar{h}, \bar{v}) \in \text{Gr}(\partial u)$ satisfying (3.3c). Feasibility $\bar{h} \in \text{dom } u$ gives $\bar{\eta} = -\Phi^* \bar{h} / \lambda \in \text{dom } R^*$, and $\bar{v} \in \partial u(\bar{h})$ is the (exact) subgradient inclusion for the pulled-back regularizer u . $\square$

Lift to a primal certificate. We now show that the proxy $(\bar{h}, \bar{v})$ of Theorem 4 lifts to a primal certificate $(\bar{\mu}, \bar{\eta})$, with $\bar{\eta} = -\Phi^* \bar{h} / \lambda$ and $\bar{v} = -\Phi \bar{\mu}$, under a mild composition constraint qualification. For a norm regularizer R the qualification holds for every bounded Φ , compact or not, and the lift is unconditional.

Proposition 3 (Pulling the subgradient back to a primal certificate). *Let $(\bar{h}, \bar{v})$ be the Brøndsted–Rockafellar proxy of Theorem 4, with $\bar{\eta} = -\Phi^* \bar{h} / \lambda$. The inclusion $\bar{v} \in \partial u(\bar{h})$ holds unconditionally. Suppose in addition the Robinson–Rockafellar constraint qualification*

$$0 \in \text{int}(\text{dom } R^* - \text{ran } \Phi^*)$$

holds or, more generally, the Attouch–Brezis condition that $\bigcup_{t \geq 0} t(\text{dom } R^ - \text{ran } \Phi^*)$ is a closed subspace of $\mathcal{B}$. Then the subdifferential chain rule holds with equality at $\bar{h}$,*

$$\partial u(\bar{h}) = -\Phi \partial R^*(\bar{\eta}) \quad \left(\text{GBL case of Remark 9: } N_{K_\lambda}(\bar{h}) = -\Phi N_{\{\|\cdot\|_\infty \leq 1\}}(\bar{\eta}) \right),$$

and $\bar{v}$ lifts to a primal certificate: there exists $\bar{\mu} \in \mathcal{B}^$ with $\bar{v} = -\Phi \bar{\mu}$ and $\bar{\eta} \in \partial R(\bar{\mu})$, a source condition for R at $\bar{\mu}$.*

Proof of Proposition 3. The inclusion $\bar{v} \in \partial u(\bar{h})$ is the proximal optimality established in the proof of Theorem 4. Re-expressing it as a source condition needs the reverse inclusion $\partial u(\bar{h}) \subseteq -\Phi \partial R^*(\bar{\eta})$, i.e. the subdifferential composition rule $\partial(g \circ A) = A^* \partial g(A \cdot)$ for $g = \lambda R^*$ and $A = -\Phi^* / \lambda$ (the inclusion $\supseteq$ is automatic). Under this constraint qualification the composition rule holds with equality and the primal minimizer is attained [Zălinescu, 2002, Section 2.8], so $\partial u(\bar{h}) = -\Phi \partial R^*(\bar{\eta})$; since $\bar{v} \in \partial u(\bar{h})$, there is $\bar{\mu} \in \partial R^*(\bar{\eta})$ with $\bar{v} = -\Phi \bar{\mu}$, and $\bar{\mu} \in \partial R^*(\bar{\eta})$ is equivalent to $\bar{\eta} \in \partial R(\bar{\mu})$ by Fenchel–Young duality. $\square$

For a norm regularizer the Robinson–Rockafellar condition holds for free, with no assumption on Φ : $\text{dom } R^* = B_{\mathcal{B}}$ is the closed unit ball of the pre-dual, so $0 \in \text{int } \text{dom } R^*$ and $R^* = \iota_{B_{\mathcal{B}}}$ is finite and continuous there; since $0 = A0 \in \text{ran } \Phi^*$, the interior condition $0 \in \text{int}(\text{dom } R^* - \text{ran } \Phi^*)$ holds at once, and a fortiori the Attouch–Brezis condition, since $\bigcup_{t \geq 0} t(\text{dom } R^* - \text{ran } \Phi^*) \supseteq \bigcup_{t \geq 0} tB_{\mathcal{B}} = \mathcal{B}$ is a closed subspace [Attouch and Brezis, 1986]. For a compact Φ with infinite-dimensional range (bandlimited or RKHS feature maps), $\text{ran } \Phi^*$ is not closed in $\mathcal{B}$, yet the interior route above still holds for a norm regularizer, so the equality is unchanged. Closedness of $\text{ran } \Phi^*$ is not on its own a constraint qualification for the chain rule: the qualification bears on the pair $(\text{dom } R^*, \text{ran } \Phi^*)$, requiring the cone $\bigcup_{t \geq 0} t(\text{dom } R^* - \text{ran } \Phi^*)$ to be a linear *subspace* and not merely closed. For $\Phi = 0$ the range $\{0\}$ is closed, yet this cone is the one generated by $\text{dom } R^*$, in general not a subspace, and the equality can fail: e.g. $R^*(f) = -\sqrt{f}$ on $[0, \infty)$, where $\partial R^*(0) = \emptyset$ while $\partial u(\bar{h}) = \{0\}$.

Consequently the proxy $\bar{v}$ therefore always arises from an actual primal measure $\bar{\mu}$ with $\bar{\eta} \in \partial R(\bar{\mu})$, for every bounded Φ . What a compact Φ can obstruct is a separate property: certificate existence in $\text{ran } \Phi^*$ for a prescribed measure μ (Remark 8, Lemma 13(iii)), not the chain rule at $\bar{h}$.

Remark 9 (Concrete proximal operators).

- (i) **Total variation (GBL).** For $R = \|\cdot\|_{\text{TV}}$ the conjugate is $R^* = \iota_{\{\|\cdot\|_\infty \leq 1\}}$, so $u = \iota_{K_\lambda}$ is the indicator of the dual-feasible set

$$K_\lambda := \{h' \in \mathcal{H} : \|\Phi^* h'\|_\infty \leq \lambda\}$$

feasible since $0 \in K_\lambda$, with $\partial u = N_{K_\lambda}$ the outward normal cone and $\text{prox}_{\tau u} = \mathcal{P}_{K_\lambda}$, the projection onto K_λ , for all $\tau > 0$.

Writing $\bar{\eta} = -\Phi^* \bar{h}/\lambda$, so that $\bar{h} \in K_\lambda$ iff $\|\bar{\eta}\|_\infty \leq 1$, the cone $N_{K_\lambda}(\bar{h})$ is carried by the saturation set $\{x \in \mathcal{X} : |\bar{\eta}(x)| = 1\}$: any measure $\bar{\mu} \in \mathcal{M}(\mathcal{X})$ supported there with $\text{sign } \bar{\mu} = \text{sign } \bar{\eta}$ (equivalently $\bar{\eta} \in \partial R(\bar{\mu})$), a dual certificate for $\bar{\mu}$ gives $-\Phi \bar{\mu} \in N_{K_\lambda}(\bar{h})$, and

$$N_{K_\lambda}(\bar{h}) = -\Phi N_{\{\|\cdot\|_\infty \leq 1\}}(\bar{\eta}),$$

with equality; the inclusion $\supseteq$ is automatic, and for a norm regularizer the reverse inclusion also holds, for every bounded Φ , since $\text{dom } R^*$ is the unit ball of the pre-dual and $0 \in \text{int dom } R^* \cap \text{ran } \Phi^*$ supplies the qualification of Proposition 3, whether or not Φ is compact.

Consequently, for any algorithmic pair (μ, h) with $\lambda L_R(\mu \| \eta) \leq \epsilon$, the proxy of Theorem 4 is an exactly dual-feasible point $\bar{h} \in K_\lambda$ with $\bar{v} \in N_{K_\lambda}(\bar{h})$, $\|\bar{h} - h\|_{\mathcal{H}} \leq \sqrt{\tau \epsilon}$ and $\|\bar{v} - (-\Phi \mu)\|_{\mathcal{H}} \leq \sqrt{\epsilon/\tau}$; its certificate $\bar{\eta} = -\Phi^* \bar{h}/\lambda$ satisfies $\|\bar{\eta}\|_\infty \leq 1$, an admissible TV certificate. The Fenchel–Young tolerance ϵ thus certifies a fully feasible TV certificate within Hilbert distance $\sqrt{\tau \epsilon}$ of the iterate.

- (ii) **Squared norm.** For $\mathcal{B}$ a Hilbert space and $R(\mu) = \frac{1}{2\lambda} \|\mu\|_{\mathcal{B}^*}^2$, we have $u(h) = \frac{1}{2} \|\Phi^* h\|_{\mathcal{B}}^2$ and $\text{prox}_{\tau u}(h) = (\text{Id} + \tau \Phi \Phi^*)^{-1} h$, the Tikhonov resolvent in $\mathcal{H}$.
- (iii) **Nuclear norm.** For $R(W) = \|W\|_*$, u is the indicator of the set $\{h : \|\Phi^* h\|_{\text{op}} \leq \lambda\}$ and $\text{prox}_{\tau u}$ is the orthogonal projection onto this set. For $\Phi = \text{Id}$ this projection clips the singular values at level λ , $\sigma_i \mapsto \min(\sigma_i, \lambda)$; it is not the singular-value soft-thresholding, which is the proximal operator of the nuclear norm itself.

Remark 10 (Cheap implementable proxy: dual rescaling). In the TV case the orthogonal projection $\mathcal{P}_{K_\lambda}$ is a continuous-domain quadratic program with infinitely many linear constraints. For numerical applications, it can be replaced by the scalar contraction

$$\tilde{h} := c h, \quad c := \min\left(1, \frac{\lambda}{\|\Phi^* h\|_\infty}\right) \in [0, 1],$$

a one-step retraction onto K_λ . Even if it is not obtained by projection, the certificate $\tilde{\eta} = -\Phi^* \tilde{h}/\lambda = c \eta$ stays in $\text{ran } \Phi^*$, and the duality-gap identity (Theorem 1) at $(\mu, \tilde{h})$ yields the computable decomposition

$$\Delta(\mu, \tilde{h}) = L_F(\Phi \mu \| \tilde{h}) + \lambda L_R(\mu \| \tilde{\eta}) \geq J_{\mathcal{B}^*}(\mu) - \inf J_{\mathcal{B}^*},$$

both pieces finite and oracle-free (see Section 5.1 for further details).

3.3 Mirror proximal refinement adapted to the fidelity

A limitation of Theorem 4 is that it is based on a prox step, hence gives a particular role to the squared Hilbert distance on $\mathcal{H}$ whereas in Section 2 we only dealt with Bregman divergences, comparing images $\Phi \mu$ through the lens of F . Also Theorem 4 is meant to obtain a point satisfying the source condition, which leaves aside the notion of mirror alignment that we introduced and discussed in Theorem 2. We now explore a Bregman variant of the previous approach, recovering the previous computations in the particular case of quadratic F .

Throughout, F is differentiable and strictly convex on $\mathcal{H}$ and F^* is differentiable on $\text{int}(\text{dom}(F^*))$, as in Section 2. For $\tau > 0$ we define the *tilted Bregman proximal step*

$$\bar{h} := \arg \min_{h' \in \mathcal{H}} \left\{ u(h') + \langle \Phi \mu, h' \rangle_{\mathcal{H}} + \frac{1}{\tau} F^*(h' | h) \right\}, \quad (3.4a)$$

and we assume that the minimum is attained at some $\bar{h}$ with $h, \bar{h} \in \text{int}(\text{dom}(F^*))$. The first-order optimality condition of (3.4a) reads

$$\bar{v} := -\Phi \mu + \frac{1}{\tau} (\nabla F^*(h) - \nabla F^*(\bar{h})) \in \partial u(\bar{h}). \quad (3.4b)$$

In the quadratic case $F = \frac{1}{2} \|\cdot - y\|_{\mathcal{H}}^2$ one has $F^*(h' | h) = \frac{1}{2} \|h' - h\|_{\mathcal{H}}^2$, since the data y only enters the linear part of F^* , so (3.4a) is $\text{prox}_{\tau u}(h - \tau \Phi \mu)$ and (3.4b) is (3.3b): the step reduces to the construction

of Theorem 4. The tilt $\langle \Phi\mu, \cdot \rangle_{\mathcal{H}}$ uses the prediction $\Phi\mu$ as a surrogate of the dual gradient $\nabla F^*(h)$, and by (2.6c) the surrogate error is the fidelity loss itself, $F(\Phi\mu \mid \nabla F^*(h)) = L_F(\Phi\mu \parallel h)$. The generalization of Carrier's proximal construction is a mirror proximal step with kernel F^* , tilted by the prediction.

The first result extends the Brøndsted–Rockafellar bounds of Theorem 4. It does not use any lift to a primal certificate.

Proposition 4 (Bregman Brøndsted–Rockafellar). *Let Assumption 1 hold and let $(\mu, h) \in \text{dom } R \times \text{dom } u$ satisfy $\lambda L_R(\mu \parallel \eta) \leq \epsilon$ with $\eta = -\Phi^*h/\lambda$. For $\tau > 0$, let $\bar{h}$ and $\bar{v} \in \partial u(\bar{h})$ be given by (3.4). Then $\bar{h} \in \text{dom } u$, so the proxy is dual-feasible, $\bar{\eta} := -\Phi^*\bar{h}/\lambda \in \text{dom } R^*$, and*

$$\frac{1}{\tau} \left[F^*(\bar{h} \mid h) + F^*(h \mid \bar{h}) \right] = \langle \bar{v} - (-\Phi\mu), h - \bar{h} \rangle_{\mathcal{H}} \leq \epsilon. \quad (3.5)$$

If moreover ∇F is M -Lipschitz then

$$\|\bar{h} - h\|_{\mathcal{H}} \leq \sqrt{M\tau\epsilon},$$

and if F is γ -strongly convex then

$$\|\bar{v} - (-\Phi\mu)\|_{\mathcal{H}} \leq \sqrt{\epsilon/(\gamma\tau)}.$$

The quadratic case, where $M = \gamma = 1$, recovers (3.3).

Proof. By Proposition 2, $L_u(h \parallel -\Phi\mu) \leq \lambda L_R(\mu \parallel \eta) \leq \epsilon$. The Fenchel–Young inequality at $\bar{h}$ gives $u^*(-\Phi\mu) \geq \langle -\Phi\mu, \bar{h} \rangle_{\mathcal{H}} - u(\bar{h})$, hence

$$\epsilon \geq u(h) + u^*(-\Phi\mu) + \langle \Phi\mu, h \rangle_{\mathcal{H}} \geq u(h) - u(\bar{h}) + \langle \Phi\mu, h - \bar{h} \rangle_{\mathcal{H}}.$$

The subgradient inequality $u(h) \geq u(\bar{h}) + \langle \bar{v}, h - \bar{h} \rangle_{\mathcal{H}}$ then yields $\epsilon \geq \langle \bar{v} + \Phi\mu, h - \bar{h} \rangle_{\mathcal{H}}$, and by (3.4b) together with the definition of the Bregman divergence,

$$\langle \bar{v} + \Phi\mu, h - \bar{h} \rangle_{\mathcal{H}} = \frac{1}{\tau} \langle \nabla F^*(h) - \nabla F^*(\bar{h}), h - \bar{h} \rangle_{\mathcal{H}} = \frac{1}{\tau} \left[F^*(\bar{h} \mid h) + F^*(h \mid \bar{h}) \right].$$

If ∇F is M -Lipschitz then F^* is $(1/M)$ -strongly convex, so the bracket is at least $\frac{1}{M} \|h - \bar{h}\|_{\mathcal{H}}^2$. If F is γ -strongly convex then ∇F^* is $(1/\gamma)$ -Lipschitz, so by co-coercivity $\langle \nabla F^*(h) - \nabla F^*(\bar{h}), h - \bar{h} \rangle_{\mathcal{H}} \geq \gamma \|\nabla F^*(h) - \nabla F^*(\bar{h})\|_{\mathcal{H}}^2 = \gamma\tau^2 \|\bar{v} + \Phi\mu\|_{\mathcal{H}}^2$. $\square$

The second result presents an exact identity for the duality gap.

Proposition 5 (Exact gap identity for the tilted step, $\tau = 1$). *In the setting of Proposition 4 with $\tau = 1$, suppose in addition that $\bar{v}$ lifts to a primal certificate:*

$$\text{there is } \bar{\mu} \in \mathcal{B}^* \text{ with } \bar{v} = -\Phi\bar{\mu} \text{ and } \bar{\eta} = -\frac{\Phi^*\bar{h}}{\lambda} \in \partial R(\bar{\mu}). \quad (3.6a)$$

Write $g := \nabla F^*(h)$, $\bar{g} := \nabla F^*(\bar{h})$ and the mirror residual $r := \Phi\mu - g$. Then the step preserves the residual,

$$\Phi\bar{\mu} - \bar{g} = \Phi\mu - g = r, \quad (3.6b)$$

and the duality gap satisfies

$$\Delta(\mu, h) - \Delta(\bar{\mu}, \bar{h}) = \lambda L_R(\mu \parallel \eta) + [F(g + r \mid g) - F(\bar{g} + r \mid \bar{g})]. \quad (3.6c)$$

In the quadratic case $F(p + r \mid p) = \frac{1}{2} \|r\|_{\mathcal{H}}^2$ does not depend on the base point p , the bracket vanishes, and (3.6c) gives $\Delta(\bar{\mu}, \bar{h}) = \Delta(\mu, h) - \lambda L_R(\mu \parallel \eta)$ for the quadratic case. Note also that by Proposition 3, the requirement (3.6a) holds under the stated constraint qualification, and unconditionally for a norm regularizer.

Proof. From (3.4b) with $\tau = 1$ and $\bar{v} = -\Phi\bar{\mu}$ we get $\Phi\bar{\mu} = \Phi\mu - g + \bar{g}$, which is (3.6b). Since $\bar{\eta} \in \partial R(\bar{\mu})$ we have $L_R(\bar{\mu} \parallel \bar{\eta}) = 0$. By the Bregman identity (2.6c), $L_F(\Phi\mu \parallel h) = F(\Phi\mu \mid \nabla F^*(h)) = F(g + r \mid g)$ and, using (3.6b), $L_F(\Phi\bar{\mu} \parallel \bar{h}) = F(\bar{g} + r \mid \bar{g})$. Subtracting the two decompositions of the gap,

$$\Delta(\mu, h) - \Delta(\bar{\mu}, \bar{h}) = L_F(\Phi\mu \parallel h) - L_F(\Phi\bar{\mu} \parallel \bar{h}) + \lambda L_R(\mu \parallel \eta),$$

gives (3.6c). $\square$

The drift term in (3.6c) is second order: it is controlled by the squared residual times the move of the mirror point, and the move is controlled by Proposition 4.

Corollary 6 (Gap bound with a second-order remainder). *In the setting of Proposition 5, assume moreover that F is γ -strongly convex and twice Gâteaux differentiable with*

$$\|\nabla^2 F(p) - \nabla^2 F(q)\|_{\text{op}} \leq L_2 \|p - q\|_{\mathcal{H}} \text{ for all } p, q \in \mathcal{H}.$$

Then

$$\Delta(\bar{\mu}, \bar{h}) \leq \Delta(\mu, h) - \lambda L_R(\mu \| \eta) + \frac{L_2}{\gamma^{3/2}} L_F(\Phi\mu \| h) \sqrt{\lambda L_R(\mu \| \eta)}. \quad (3.7)$$

No tolerance parameter appears: the bound uses Proposition 4 with the choice $\epsilon = \lambda L_R(\mu \| \eta)$, which is admissible since ϵ need not be small, and every term on the right-hand side is computable from the pair (μ, h) . The remainder vanishes when F is quadratic, where $L_2 = 0$: the proxy then improves the gap by exactly $\lambda L_R(\mu \| \eta)$.

Proof. Set $\psi(p) := F(p + r | p) = F(p + r) - F(p) - \langle \nabla F(p), r \rangle_{\mathcal{H}}$. For every $p \in \mathcal{H}$ and every direction $w \in \mathcal{H}$, using $\nabla F(p + r) - \nabla F(p) = \int_0^1 \nabla^2 F(p + sr) r \, ds$ and the self-adjointness of $\nabla^2 F(p)$,

$$d\psi(p)[w] = \left\langle \int_0^1 [\nabla^2 F(p + sr) - \nabla^2 F(p)] r \, ds, w \right\rangle_{\mathcal{H}},$$

so $\|d\psi(p)\| \leq \int_0^1 L_2 s \|r\|_{\mathcal{H}}^2 \, ds = \frac{L_2}{2} \|r\|_{\mathcal{H}}^2$ and, by the mean value inequality, $|\psi(g) - \psi(\bar{g})| \leq \frac{L_2}{2} \|r\|_{\mathcal{H}}^2 \|g - \bar{g}\|_{\mathcal{H}}$. By γ -strong convexity, $L_F(\Phi\mu \| h) = F(g + r | g) \geq \frac{\gamma}{2} \|r\|_{\mathcal{H}}^2$, and Proposition 4 with $\tau = 1$ and $\epsilon = \lambda L_R(\mu \| \eta)$ gives $\|g - \bar{g}\|_{\mathcal{H}} = \|\bar{v} + \Phi\mu\|_{\mathcal{H}} \leq \sqrt{\lambda L_R(\mu \| \eta)}/\gamma$. Hence the bracket in (3.6c) is at least $-\frac{L_2}{2} \cdot \frac{2}{\gamma} L_F(\Phi\mu \| h) \cdot \sqrt{\lambda L_R(\mu \| \eta)}/\gamma$, which is (3.7). $\square$

Corollary 7 (When the tilted step improves the gap). *Write $\delta := L_F(\Phi\bar{\mu} \| \bar{h}) - L_F(\Phi\mu \| h)$ for the change of fidelity loss and $\kappa := \frac{1}{\tau} [F^*(\bar{h} | h) + F^*(h | \bar{h})] \geq 0$. Since $L_R(\bar{\mu} \| \bar{\eta}) = 0$, the gap decomposition gives, for every $\tau > 0$,*

$$\Delta(\mu, h) - \Delta(\bar{\mu}, \bar{h}) = \lambda L_R(\mu \| \eta) - \delta, \quad (3.8)$$

so the tilted step improves the gap if and only if $\delta \leq \lambda L_R(\mu \| \eta)$. Two sufficient regimes make this explicit.

- (a) **Step-size balance.** By Proposition 4 with $\epsilon = \lambda L_R(\mu \| \eta)$ one has $\kappa \leq \lambda L_R(\mu \| \eta)$ unconditionally, so $\delta \leq \kappa$ suffices, with margin $\lambda L_R(\mu \| \eta) - \delta \geq \kappa - \delta \geq 0$.
- (b) **Near-quadratic, $\tau = 1$.** Under the hypotheses of Corollary 6, improvement holds whenever $\frac{L_2}{\gamma^{3/2}} L_F(\Phi\mu \| h) \leq \sqrt{\lambda L_R(\mu \| \eta)}$, and always for quadratic F ($L_2 = 0$), where the improvement equals $\lambda L_R(\mu \| \eta)$.

Remark 11 (Scope of the tilted step).

- (i) **Free parameter τ .** For $\tau \neq 1$ the residual is not preserved: $\Phi\bar{\mu} - \bar{g} = r + (1 - \frac{1}{\tau})(g - \bar{g})$, and (3.6c) picks up an extra first order term. Proposition 4 keeps the free parameter τ , as in Theorem 4.
- (ii) **Local constants.** For the losses of Section 4, the constants γ and L_2 are finite on sublevel sets only in general; in the Kullback–Leibler and entropic cases L_2 blows up at the boundary of $\text{dom } F$, so (3.7) should be applied on a sublevel set containing the iterates.

4 Applications: specific fidelities and regularizers

In this section we apply the F+R framework to a catalogue of popular data-fidelity terms and regularizers, deriving for each loss its Fenchel–Young geometry (primal F , dual F^* , and discrepancy L_F) and for each penalty the triple (R, R^*, L_R) , so that the duality gap $\Delta(\mu, h) = L_F(\Phi\mu \| h) + \lambda L_R(\mu \| \eta)$ of Theorem 1 becomes explicit on each problem class. Three of our results hold for *every* data-fidelity term below: the early-stopping certification of Theorem 4, whose hypotheses (Assumption 1) constrain only the

regularizer R ; this gap decomposition, in the general form of Theorem 1 since several of these losses leave $\mathcal{F}_0$; and the Fenchel–Young error bounds of Theorem 3, which need only a proper, l.s.c. convex F , a source-condition certificate, and $J_{\mathcal{B}^*}(\mu^*) < +\infty$ (that is $\Phi\mu^* \in \text{dom } F$, automatic when $\text{dom } F = \mathcal{H}$).

The strictly-convex $\mathcal{C}^1$ class $\mathcal{F}_0$ is needed only for the Mirror-Alignment characterization, the uniqueness of the aligned certificate, and, through the weak-coercivity bound (2.2), the existence of a minimizer. Least squares, Poisson with positive counts $y > 0$, and the data-coupled iOT primal (4.1d) with $\pi^{\text{obs}} > 0$ lie in $\mathcal{F}_0$; the logistic, Huber, squared-hinge, and Kullback–Leibler do not, and we record each loss’s $\mathcal{F}_0$ status in the loss-by-loss computations of Appendix B. The exact-recovery analysis built on these geometries is developed in the companion paper [Aubin-Frankowski and De Castro, 2026].

4.1 Specific data-fidelity terms

4.1.1 Common losses and their Fenchel conjugates

Each common data-fidelity term is separable, $F(h) = \sum_{i=1}^n f(h_i)$, and likewise $F^*(w) = \sum_i f^*(w_i)$ and $L_F(h \parallel w) = \sum_i \ell_F(h_i, w_i)$, with y_i the paired observation. Table 1 collects the scalar per-coordinate summands for six continuously differentiable ($\mathcal{C}^1$) losses (least squares, logistic, Poisson, Kullback–Leibler, Huber, squared-hinge); the full derivations and the scope condition of each loss are in Appendix B.

Loss	$f(h)$	$f^*(w)$	$\ell_F(h, w)$	$\mathcal{F}_0$
Least squares	$\frac{1}{2}(h - y)^2$	$\frac{1}{2}w^2 + wy$ $(-yw) \log(-yw)$	$\frac{1}{2}(h - w - y)^2$	Yes
Logistic	$\log(1 + e^{-yh})$	$+(1 + yw) \log(1 + yw)$	$f(h) + f^*(w) - wh$	No
Poisson	$e^h - yh$	$(w + y) \log(w + y) - (w + y)$	$D_{\text{KL}}(w + y \parallel e^h)$	Yes ($y > 0$)
Kullback–Leibler	$h \log \frac{h}{y} - h + y$	$y(e^w - 1)$	$D_{\text{KL}}(h \parallel ye^w)$	No
Huber	$\rho_\delta(h - y)$	$\frac{1}{2}w^2 + wy$	$f(h) + f^*(w) - wh$	No
Squared hinge	$\frac{1}{2} \max(0, 1 - yh)^2$	$\frac{1}{2}w^2 + yw$	$f(h) + f^*(w) - wh$	No

Table 1: Common continuously differentiable ($\mathcal{C}^1$) data-fidelity losses and their Fenchel conjugates, as scalar per-coordinate summands ($F(h) = \sum_i f(h_i)$, and likewise for F^* and L_F , with y_i the paired observation). In every row $\ell_F(h, w) = f(h) + f^*(w) - hw$; the closed forms shown for least squares, Poisson and Kullback–Leibler follow by simplification. The conjugate f^* is finite only on $-yw \in [0, 1]$ (logistic), $w + y \geq 0$ (Poisson), $|w| \leq \delta$ (Huber) and $-yw \geq 0$ (squared hinge), and is $+\infty$ otherwise; $\rho_\delta(r) = \frac{1}{2}r^2$ for $|r| \leq \delta$ and $\delta(|r| - \frac{1}{2}\delta)$ otherwise; Poisson requires $y_i \geq 0$ and Kullback–Leibler requires $y_i > 0$, and the Kullback–Leibler primal f (hence its Fenchel–Young loss ℓ_F) is defined only on $h \geq 0$, with the convention $0 \log 0 = 0$; the logistic and squared-hinge rows assume binary labels $y_i \in \{-1, +1\}$, under which the tabulated conjugates hold. The last column records membership in the class $\mathcal{F}_0$ of Definition 3 (finite-valued on $\mathcal{H}$, $\mathcal{C}^1$, strictly convex, and weakly coercive (2.2)).

4.1.2 Inverse optimal transport (iOT)

In the iOT setting, following Andrade et al. [2024, Section 2], the goal is to learn a cost matrix $C \in \mathbb{R}^{n \times n}$ from an observed (unbalanced) transport plan $\pi^{\text{obs}} \in \mathbb{R}_+^{n \times n}$.

Inverse optimal transport. Given $\mathbb{X}$ and $\mathbb{Y}$ two compact subsets of $\mathbb{R}^d$, a continuous cost $c : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{R}$, probability measures $\alpha \in \mathcal{P}(\mathbb{X})$, $\beta \in \mathcal{P}(\mathbb{Y})$, and the set $\Pi(\alpha, \beta) \subset \mathcal{P}(\mathbb{X} \times \mathbb{Y})$ of couplings with marginals α and β , the problem of entropic optimal transport is to find $\pi^* \in \arg \min_{\pi \in \Pi(\alpha, \beta)} \langle c, \pi \rangle + \varepsilon D_{\text{KL}}(\pi | \alpha \otimes \beta)$. Denoting by $\Omega(\pi) := \varepsilon D_{\text{KL}}(\pi | \alpha \otimes \beta) + \iota_{\{\pi \in \Pi(\alpha, \beta)\}}(\pi)$, first-order conditions give that $-c \in \partial\Omega(\pi^*)$. This suggests to introduce the corresponding Fenchel–Young loss. However in iOT, instead of looking for π^* given a c , we look for c given a π^{obs} . This leads to considering

$$F_{\text{iOT}}(c) := L_\Omega(-c \parallel \pi^{\text{obs}}) = \Omega^*(-c) + \langle c, \pi^{\text{obs}} \rangle + \Omega(\pi^{\text{obs}}). \quad (4.1a)$$

Known duality computations on $\Omega^*(-c)$ give that

$$\begin{aligned}\Omega^*(-c) &= \sup_{\pi \in \Pi(\alpha, \beta)} \langle -c, \pi \rangle - \varepsilon D_{\text{KL}}(\pi | \alpha \otimes \beta) \\ &= - \sup_{f \in \mathcal{C}(\mathbb{X}), g \in \mathcal{C}(\mathbb{Y})} \left\{ \langle f, \alpha \rangle + \langle g, \beta \rangle - \varepsilon \int \left(\exp \left(\frac{f + g - c}{\varepsilon} \right) - 1 \right) d\alpha(x) d\beta(y) \right\}\end{aligned}\quad (4.1b)$$

Note that $\pi^* \in \arg \min \{ \langle c, \pi \rangle + \varepsilon D_{\text{KL}}(\pi | \alpha \otimes \beta) : \pi \in \Pi(\alpha, \beta) \}$ is invariant, due to the marginals constraints, when replacing c by $\tilde{c} = c - f - g$ for any f, g . Hence we can arbitrarily consider a reduced problem defined on $\tilde{c}$, that is equivalent to (4.1a) when dropping the constant term $\Omega(\pi^{\text{obs}})$,

$$F_{\text{red.iOT}}(\tilde{c}) := \langle \tilde{c}, \pi^{\text{obs}} \rangle + \varepsilon \int \exp \left(\frac{-\tilde{c}}{\varepsilon} \right) d\alpha(x) d\beta(y) \quad (4.1c)$$

Instead of α and β , one has in practice samples $\hat{\alpha} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\hat{\beta} = \frac{1}{m} \sum_{j=1}^m \delta_{y_j}$ deduced from the marginals of the discrete π^{obs} . Unbalanced settings would appear from handling the mismatch between (α, β) and $(\hat{\alpha}, \hat{\beta})$. For simplicity we assume the latter coincide and can then write the problem in matrix form $C_{i,j} = \tilde{c}(x_i, y_j)$ and with functional

$$F(C) := \frac{\varepsilon}{mn} \sum_{i,j} \exp \left(-\frac{C_{i,j}}{\varepsilon} \right) + \langle C, \pi^{\text{obs}} \rangle. \quad (4.1d)$$

where the key term is given by

$$F_{\text{gen}}(C) = \varepsilon \sum_{i,j} \exp \left(-\frac{C_{i,j}}{\varepsilon} \right), \quad (4.1e)$$

The gradient $\nabla F(C) = -\pi_\varepsilon(C) + \pi^{\text{obs}}$, for $[\pi_\varepsilon(C)]_{i,j} := \frac{1}{mn} e^{-C_{i,j}/\varepsilon} \geq 0$, vanishes precisely at $C = C^{\text{obs}} := -\varepsilon \log \pi^{\text{obs}}$ when $\pi^{\text{obs}} > 0$ entrywise. The dual computation, Fenchel–Young loss, and local sharpening below are written for F_{gen} for concision; the corresponding objects for F differ by a known affine shift in the dual variable ($p \mapsto p - \pi^{\text{obs}}$) which does not affect Hessians, Fenchel–Young loss values along the regularization path, or the certificate analysis. Similarly we incorporate the term mn into the definition of C .

Dual data-fidelity and gradient. The Fenchel conjugate $F_{\text{gen}}^* : \mathbb{R}^{n \times n} \rightarrow \mathbb{R} \cup \{+\infty\}$ is the negative generalized Boltzmann–Shannon entropy:

$$F_{\text{gen}}^*(\pi) = \begin{cases} \varepsilon \sum_{i,j} (-\pi_{i,j}) (\log(-\pi_{i,j}) - 1) & \text{if } \pi \leq 0, \\ +\infty & \text{otherwise.} \end{cases} \quad (4.1f)$$

The gradient of the dual maps a transport plan back to the cost that generated it:

$$[\nabla F_{\text{gen}}^*(\pi)]_{i,j} = -\varepsilon \log(-\pi_{i,j}). \quad (4.1g)$$

This confirms the inverse relationship: if $\pi = -\exp(-C/\varepsilon)$, then $C = -\varepsilon \log(-\pi)$.

Fenchel–Young loss. The Fenchel–Young loss $L_F(C \parallel -\pi^{\text{obs}})$ corresponds to the generalized Kullback–Leibler divergence:

$$L_{F_{\text{gen}}}(C \parallel -\pi^{\text{obs}}) = \varepsilon D_{\text{KL}}(\pi^{\text{obs}} \parallel \pi_\varepsilon(C)) = \varepsilon \sum_{i,j} \left(\pi_{i,j}^{\text{obs}} \log \left(\frac{\pi_{i,j}^{\text{obs}}}{e^{-C_{i,j}/\varepsilon}} \right) - \pi_{i,j}^{\text{obs}} + e^{-C_{i,j}/\varepsilon} \right). \quad (4.1h)$$

4.2 Popular regularizers and their Fenchel–Young geometry

We now consider the regularizer side R and record, for a catalogue of standard sparsity-promoting penalties, the regularizer R , its Fenchel conjugate R^* , and the associated Fenchel–Young loss L_R . We order them so that the canonical instance, the total-variation norm of signed measures (the GBL regularizer), comes first, followed by its vector-valued extension to group/joint TV, then *analysis sparsity* (the regularizer acting on a linear transform $\mu = Lu$, subsuming TV-of-gradient as $L = \nabla$) and the *nuclear norm* on low-rank operators.

4.2.1 Regularization by the total variation norm (sparse spikes)

We consider the classical problem of recovering a sparse signed measure $\mu \in \mathcal{M}(\mathcal{X})$ (a sum of Dirac masses) from measurements. The regularizer is the total variation norm (or Radon norm).

Primal and dual functions. The regularizer $R : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{R}$ is the total variation norm:

$$R(\mu) = \|\mu\|_{\text{TV}} := \sup \left\{ \int_{\mathcal{X}} f(x) d\mu(x) \mid f \in \mathcal{C}_0(\mathcal{X}), \|f\|_{\infty} \leq 1 \right\}. \quad (4.2a)$$

The pre-dual space is $\mathcal{C}_0(\mathcal{X})$ equipped with the supremum norm. The Fenchel conjugate R^* is the indicator function of the unit ball of continuous functions:

$$R^*(f) = \iota_{\{\|f\|_{\infty} \leq 1\}}(f) = \begin{cases} 0 & \text{if } \sup_{x \in \mathcal{X}} |f(x)| \leq 1, \\ +\infty & \text{otherwise} \end{cases} \quad (4.2b)$$

Fenchel–Young loss. The associated Fenchel–Young loss penalizes the lack of correlation between the measure and the dual certificate (or can be seen as a slack in Hölder’s inequality), while enforcing the dual constraint:

$$L_R(\mu \parallel f) = \begin{cases} \|\mu\|_{\text{TV}} - \langle f, \mu \rangle_{\mathcal{C}, \mathcal{M}} & \text{if } \|f\|_{\infty} \leq 1, \\ +\infty & \text{otherwise} \end{cases} \quad (4.2c)$$

This loss is zero if and only if f is a subgradient of the TV norm at μ , meaning f saturates the norm constraints on the support of μ with the correct sign.

4.2.2 Regularization by group sparsity (vector-valued measures)

We consider the recovery of a vector-valued measure $\mu \in \mathcal{M}(\mathcal{X}; \mathbb{R}^d)$ where the channels are coupled. The goal is to recover spikes that share the same support across all d dimensions (joint sparsity).

Primal and dual functions. The regularizer R is the total variation of the vector measure, defined using the Euclidean norm $|\cdot|_2$ on $\mathbb{R}^d$ (acting as a group-Lasso penalty in the continuum):

$$R(\mu) = \|\mu\|_{\mathcal{M}, 2} := \int_{\mathcal{X}} \left| \frac{d\mu}{d|\mu|}(x) \right|_2 d|\mu|(x), \quad (4.3a)$$

where $\frac{d\mu}{d|\mu|}$ is the Radon–Nikodým derivative (the local orientation vector) and $|\mu|$ denotes the total-variation (Euclidean) measure of the vector-valued μ . The pre-dual space is $\mathcal{C}(\mathcal{X}; \mathbb{R}^d)$, paired with $\mathcal{M}(\mathcal{X}; \mathbb{R}^d)$ via $\langle f, \mu \rangle_{\mathcal{C}, \mathcal{M}} = \int_{\mathcal{X}} \langle f(x), d\mu(x) \rangle_{\mathbb{R}^d}$. The Fenchel conjugate R^* is the indicator of the unit ball of the dual norm (the sup-norm of the pointwise Euclidean norms):

$$R^*(f) = \iota_{K_R}(f) \quad \text{with} \quad K_R := \left\{ f \in \mathcal{C}(\mathcal{X}; \mathbb{R}^d) \mid \sup_{x \in \mathcal{X}} |f(x)|_2 \leq 1 \right\}. \quad (4.3b)$$

Fenchel–Young loss. The associated Fenchel–Young loss enforces the “block-constraint” on the dual certificate:

$$L_R(\mu \parallel f) = \begin{cases} \|\mu\|_{\mathcal{M}, 2} - \langle f, \mu \rangle_{\mathcal{C}, \mathcal{M}} & \text{if } f \in K_R, \\ +\infty & \text{otherwise} \end{cases} \quad (4.3c)$$

This loss is zero if and only if $f \in \partial R(\mu)$, i.e. $|f(x)|_2 = 1$ and $f(x)$ is aligned with $\frac{d\mu}{d|\mu|}(x)$ on $\text{supp } \mu$.

4.2.3 Sparse analysis prior

We consider the recovery of a signal $u \in \mathcal{U}$ (a Banach signal space, e.g. $L^2(\mathbb{R}^d)$ or a Sobolev space) by imposing sparsity on its analysis transform $\mu = Lu$, where $L : \mathcal{U} \rightarrow \mathcal{M}(\mathcal{X})$ is a bounded linear operator (the analysis operator). Following the analysis-sparsity representer theorem of [Boyer et al. \[2019\]](#), this is equivalent to optimizing over the measure μ with a structural constraint. The special case $L = \nabla$ (TV regularization of a gradient) admits a simple Meyer-norm dual and is treated separately in [Section 4.2.4](#).

Standing assumption. Throughout this subsection we assume $\text{Im}(L)$ is *weak-* closed* in $\mathcal{M}(\mathcal{X})$ (equivalently $\text{Im}(L) = \text{Ker}(L^*)^\perp$, the annihilator of $\text{Ker}(L^*)$ in $\mathcal{M}(\mathcal{X})$), so that the regularizer R below is weak-* l.s.c. and $R = R^{**}$.

Primal and dual functions. The regularizer $R : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{R} \cup \{+\infty\}$ combines the total variation norm with the hard constraint that the measure must belong to the (weak-* closed) image of the analysis operator:

$$R(\mu) = \|\mu\|_{\text{TV}} + \iota_{\text{Im}(L)}(\mu) = \begin{cases} \|\mu\|_{\text{TV}} & \text{if } \exists u \in \mathcal{U}, \mu = Lu, \\ +\infty & \text{otherwise.} \end{cases} \quad (4.4a)$$

The dual function R^* is the indicator of a quotient-norm ball: a dual vector $f \in \mathcal{C}_0(\mathcal{X})$ is feasible when its sup-norm distance to $\text{Ker}(L^*)$, the kernel of the pre-adjoint operator $L^* : \mathcal{C}(\mathcal{X}) \rightarrow \mathcal{U}^*$, is at most one,

$$R^*(f) = \iota_{\{\text{dist}_\infty(\cdot, \text{Ker}(L^*)) \leq 1\}}(f), \quad \text{dist}_\infty(f, \text{Ker}(L^*)) := \inf_{g \in \text{Ker}(L^*)} \|f - g\|_\infty, \quad (4.4b)$$

i.e. R^* is the indicator of the unit ball of the quotient norm on $\mathcal{C}_0(\mathcal{X})/\text{Ker}(L^*)$; the standing assumption ($\text{Im}(L)$ weak-* closed) identifies $\text{Im}(L)$ with the dual of this quotient space, which gives the formula. The exact infimal convolution $\inf_{g \in \text{Ker}(L^*)} \iota_{\{\|\cdot\|_\infty \leq 1\}}(f - g)$ takes the same value when the distance is attained; in general the conjugate is its lower semi-continuous closure, because the distance to a closed subspace of $\mathcal{C}(\mathcal{X})$ need not be attained.

Fenchel–Young loss. The Fenchel–Young loss L_R penalizes the lack of an “analysis sparsity” certificate. For a measure $\mu \in \text{Im}(L)$ and a dual certificate f :

$$L_R(\mu \parallel f) = \begin{cases} \|\mu\|_{\text{TV}} - \langle f, \mu \rangle_{\mathcal{C}, \mathcal{M}} & \text{if } \mu \in \text{Im}(L) \text{ and } \text{dist}_\infty(f, \text{Ker}(L^*)) \leq 1, \\ +\infty & \text{otherwise.} \end{cases} \quad (4.4c)$$

The loss is finite only when $\text{dist}_\infty(f, \text{Ker}(L^*)) \leq 1$, i.e. when f is close to a function bounded by one, up to a component $g \in \text{Ker}(L^*)$ invisible to the signal space.

4.2.4 Regularization by the total variation norm of the gradient

Throughout this subsection, $\mathcal{X} \subset \mathbb{R}^d$ is a *bounded Lipschitz domain*. It treats the special case $L = \nabla$ of analysis sparsity (Section 4.2.3): the regularizer is the TV norm of a gradient field $\mu = \nabla u$ of a function of bounded variation. We keep it as a separate subsubsection because the Meyer-norm structure of the dual needs its own derivation, as it is specific to the divergence operator $L^* = -\text{div}$ (the sign is immaterial here, as the dual constraint set K_G below is symmetric under $f \mapsto -f$).

Primal and dual functions. Let $\mathcal{B}^* = \mathcal{M}(\mathcal{X}; \mathbb{R}^d)$ be the space of vector-valued Radon measures. We consider the regularization of the gradient of a function u of bounded variation, writing

$$BV_0(\mathcal{X}) := \{u \in BV(\mathbb{R}^d) : u = 0 \text{ a.e. on } \mathbb{R}^d \setminus \mathcal{X}\}$$

for the BV functions vanishing outside $\mathcal{X}$. The regularizer $R : \mathcal{M}(\mathcal{X}; \mathbb{R}^d) \rightarrow \mathbb{R} \cup \{+\infty\}$ is defined as the total variation norm restricted to gradient fields:

$$R(\mu) = \begin{cases} \|\mu\|_{\text{TV}} & \text{if } \mu = \nabla u \text{ for some } u \in BV_0(\mathcal{X}), \\ +\infty & \text{otherwise} \end{cases} \quad (4.5a)$$

where the equality $\mu = \nabla u$ holds in the distributional sense. The constraint set $\{\nabla u : u \in BV_0(\mathcal{X})\}$ is *weak-* closed* in $\mathcal{M}(\mathcal{X}; \mathbb{R}^d)$: if $\nabla u_n \rightharpoonup^* \mu$ with $u_n \in BV_0(\mathcal{X})$, then, assuming $\mathcal{X}$ is a bounded Lipschitz domain in $\mathbb{R}^d$, the Poincaré inequality bounds (u_n) in BV , the compact embedding $BV \hookrightarrow L^1$ gives a limit $u \in BV_0(\mathcal{X})$ (identifying u with its extension by zero to $\mathbb{R}^d$), and $\nabla u = \mu$. Hence R is weak-* l.s.c. and the standing assumption of Section 4.2.3 holds here. The pre-dual space is $\mathcal{B} = \mathcal{C}(\mathcal{X}; \mathbb{R}^d)$. Following

the framework of De Castro et al. [2023a], the Fenchel conjugate R^* is the indicator function of the convex set K_G defined by a constraint on the divergence:

$$R^*(f) = \iota_{K_G}(f) \quad \text{with} \quad K_G := \left\{ f \in \mathcal{C}(\mathcal{X}; \mathbb{R}^d) \mid \left| \int_E \operatorname{div}(f) \right| \leq \operatorname{Per}(E), \quad \forall E \subset \mathcal{X} \right\}, \quad (4.5b)$$

where E ranges over sets of finite perimeter in $\mathcal{X}$. This condition is equivalent to $\|\operatorname{div}(f)\|_G \leq 1$, where $\|\cdot\|_G$ denotes the Meyer norm (the dual of the BV semi-norm). The Meyer norm $\|\cdot\|_G$, introduced in Meyer [2001], is defined on the space of distributions G that can be represented as the divergence of a bounded vector field. Formally:

$$\|v\|_G := \inf \{ \|g\|_{L^\infty(\mathcal{X}; \mathbb{R}^d)} \mid v = \operatorname{div}(g), g \in L^\infty(\mathcal{X}; \mathbb{R}^d) \}. \quad (4.5c)$$

Equivalently, the identity $\|v\|_G = \sup\{\langle u, v \rangle : u \in BV_0(\mathcal{X}), \|\nabla u\|_{\operatorname{TV}} \leq 1\}$ exhibits $\|\cdot\|_G$ as the dual norm of the BV semi-norm (consistent with the description above).

Fenchel–Young loss. The associated Fenchel–Young loss reads:

$$L_R(\mu \parallel f) = \begin{cases} \|\mu\|_{\operatorname{TV}} - \langle f, \mu \rangle_{\mathcal{C}, \mathcal{M}} & \text{if } \mu = \nabla u \text{ for some } u \in BV_0(\mathcal{X}) \text{ and } f \in K_G, \\ +\infty & \text{otherwise} \end{cases} \quad (4.5d)$$

Stability under primal non-uniqueness. The regularizer R is the TV norm restricted to gradient fields, hence positively 1-homogeneous and not strictly convex, so primal uniqueness is generally not guaranteed in the Meyer-norm setting. The total duality gap nevertheless decomposes as

$$\Delta(\mu, h) = \underbrace{L_F(\Phi\mu \parallel h)}_{\text{prediction-space control}} + \lambda \underbrace{L_R(\mu \parallel \eta)}_{\text{source-condition control}}. \quad (4.5e)$$

For any μ with $J_{\mathcal{B}^*}(\mu) \leq J_{\mathcal{B}^*}(\mu^*)$, the prediction bound $L_F(\Phi\mu \parallel h^*) \leq L_F(\Phi\mu^* \parallel h^*)$ (Theorem 3) and Carlier’s sharpened Fenchel–Young inequality [Carlier, 2023, Lemma 1.1] applied at $(\Phi\mu, h^*)$ imply that the squared Hilbert distance of $(\Phi\mu, \tau h^*)$ to the graph of $\tau\partial F$ is at most $2\tau L_F(\Phi\mu^* \parallel h^*)$ (cf. Theorem 3 (2.13e)); the prediction $\Phi\mu$ is thus tethered to the optimal one even when the primal measure is non-unique. Beyond this static control, the algorithmic stability of Section 3 (Theorem 4) supplies an exactly feasible dual proxy for any iterate with $\lambda L_R(\mu \parallel \eta) \leq \epsilon$, so the Meyer-norm setting also enjoys the early-stopping guarantee.

4.2.5 Regularization by nuclear norm

We consider low-rank matrix recovery, where the optimization variable is a matrix $\mu \in \mathbb{R}^{n \times m}$ (identified here as the measure). The regularizer is the nuclear norm (or Schatten 1-norm), which promotes low-rank solutions.

Primal and dual functions. Let $R : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ be the nuclear norm defined by the sum of the singular values:

$$R(\mu) := \|\mu\|_* = \sum_{i=1}^{\min(n, m)} \sigma_i(\mu). \quad (4.6a)$$

The pre-dual space is $\mathbb{R}^{n \times m}$ itself, self-dual under the trace inner product $\langle f, \mu \rangle = \operatorname{tr}(f^\top \mu)$. The dual norm is the operator norm (spectral norm), denoted $\|\cdot\|_{\operatorname{op}}$ (the largest singular value). The Fenchel conjugate R^* is the indicator function of the spectral unit ball:

$$R^*(f) = \iota_{\{\|\cdot\|_{\operatorname{op}} \leq 1\}}(f) = \begin{cases} 0 & \text{if } \|f\|_{\operatorname{op}} \leq 1, \\ +\infty & \text{otherwise.} \end{cases} \quad (4.6b)$$

Fenchel–Young loss. The associated Fenchel–Young loss L_R measures the gap between the nuclear norm and the linear pairing with the dual variable. It enforces the spectral constraint $\|f\|_{\text{op}} \leq 1$ as a hard (indicator) constraint on the dual certificate:

$$L_R(\mu \parallel f) = \begin{cases} \|\mu\|_* - \langle f, \mu \rangle & \text{if } \|f\|_{\text{op}} \leq 1, \\ +\infty & \text{otherwise.} \end{cases} \quad (4.6c)$$

This loss is zero if and only if f is a subgradient of the nuclear norm at μ : for a reduced SVD $\mu = U \text{diag}(\sigma) V^\top$, this holds if and only if $f = UV^\top + W$ with $\|W\|_{\text{op}} \leq 1$, $U^\top W = 0$ and $WV = 0$.

5 Numerical experiments

We focus on the Fenchel–Young gap $\Delta = L_F + \lambda L_R$ of Theorem 1. Throughout this section, J denotes the primal objective $J_{\mathcal{B}^*}$ of (2.3) (and its finite-dimensional instances). Section 5.1 runs on a synthetic 1D Gaussian super-resolution BLASSO problem and tracks Δ_k along the Conic Particle Gradient Descent (CPGD) of Chizat [2022], De Castro et al. [2023b].

This multiplicative amplitude update is a mirror-descent step for the total-variation regularizer of Section 2. The dual certificate η_k converges to the source condition while the inset gap Δ_k decreases with no oracle, and the decomposition of Δ_k tracks the true optimality gap (and, as the theory guarantees, dominates it), which is the operational early-stopping signal. Section 5.2 applies the same scalar diagnostic to the deep-learning optimizers Lion-K and Muon, certifying that in proximal form they solve the $F + \lambda R$ program, the gap Δ_t decreasing to 0 along the trajectory. Throughout, the single scalar Δ bounds the distance to optimality without knowing the optimum. An open-source notebook reproduces the figures below in full.²

We call Mirror-Alignment dual certificate (at iterate μ_k) the continuous function η_k defined by

$$\eta_k = -\frac{\Phi^* \nabla F(\Phi \mu_k)}{\lambda}.$$

We call the *rescaled* Mirror-Alignment dual the continuous function $\tilde{\eta}_k$, which satisfies dual feasibility $\tilde{h}_k \in K_\lambda$, defined by

$$\begin{aligned} \tilde{h}_k &:= c_k h_k, \quad h_k = \nabla F(\Phi \mu_k), \quad \tilde{\eta}_k = -\Phi^* \tilde{h}_k / \lambda = c_k \eta_k, \\ c_k &:= \min\left(1, \frac{\lambda}{\|\Phi^* h_k\|_\infty}\right) = \min\left(1, \frac{1}{\|\eta_k\|_\infty}\right) \in [0, 1]; \\ K_\lambda &:= \left\{h' \in \mathcal{H} : \|\Phi^* h'\|_\infty \leq \lambda\right\}, \quad \tilde{h}_k \in K_\lambda; \end{aligned} \quad (5.1a)$$

and the duality gap of the BLASSO program

$$\Delta(\mu, h) = L_F(\Phi \mu \parallel h) + \lambda L_R(\mu \parallel \eta), \quad \eta = -\Phi^* h / \lambda, \quad (5.1b)$$

which decomposes into a data-fidelity Fenchel–Young loss L_F and a regularizer Fenchel–Young loss L_R .

5.1 Dual certificate evolution along CPGD

Figure 2 tracks the dual certificate η_k along the CPGD trajectory of a synthetic 1D super-resolution problem: four spikes at (0.15, 0.31, 0.50, 0.83) with signed amplitudes (+1.4, −0.6, +0.9, −0.7), Gaussian feature map of bandwidth $\sigma_\varphi = 0.04$, $N = 200$ samples, observation noise $\sigma_w = 10^{-2}$, $\lambda = 10^{-2}$, and $T = 30\,000$ iterations, the particles being initialized on a uniform grid of $[0, 1]$ with alternating signs. The convergence pattern visible in Figure 2 is the operational form of the early-stopping theory: the rescaled dual $\tilde{h}_k \in K_\lambda$ is dual-feasible at every iteration (Remark 10), and the Fenchel–Young duality gap $\Delta_k = L_F(\Phi \mu_k \parallel \tilde{h}_k) + \lambda L_R(\mu_k \parallel \tilde{\eta}_k)$ is an oracle-free optimality certificate (Theorem 1 and (2.9)) computed without any access to the unknown μ^* ; along the trajectory it is observed to decrease to zero.

In Figure 2, the inset reports the oracle-free gap Δ_k . Across the four snapshots Δ_k falls by more than four orders of magnitude ($1.8 \times 10^{-1} \rightarrow 1.2 \times 10^{-5}$), $c_k \rightarrow 1$, and $\|\eta_k\|_\infty \rightarrow 1$ with on-support saturation

²Code and companion notebook available at <https://github.com/ydecastro/fenchel-young-gaps>.

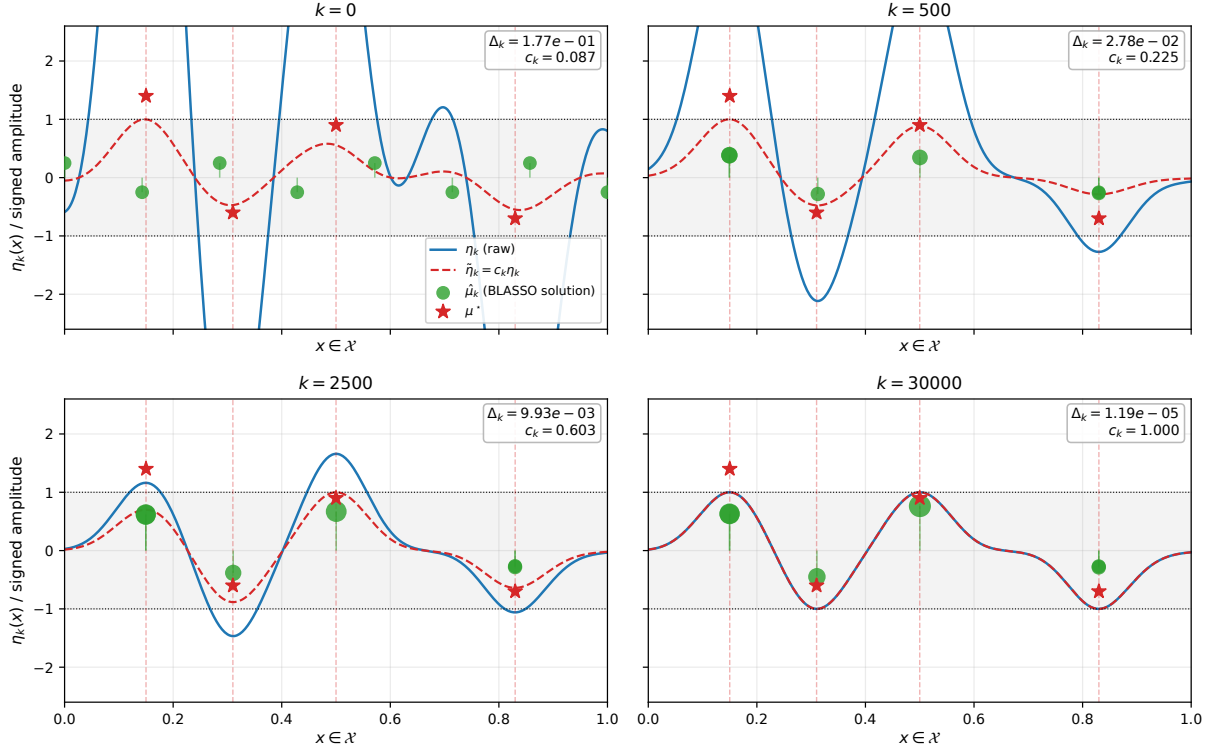


Figure 2: Dual certificate evolution along CPGD (synthetic 1D Gaussian). Each panel overlays, at iteration k : the Mirror-Alignment certificate $\eta_k = -\Phi^* h_k / \lambda$ (blue); its rescaled contraction $\tilde{\eta}_k = c_k \eta_k$, $c_k = \min(1, 1/\|\eta_k\|_\infty)$ (red dashed); the recovered BLASSO solution $\hat{\mu}_k$ as individual particles (green; one stem per active CPGD particle, marker size scaled by amplitude); and the truth μ^* (red stars).

$\eta_k(\hat{x}_i) \approx \text{sign}(\hat{a}_i)$ at the active atoms $\hat{a}_i \delta_{\hat{x}_i}$ of μ_k , the operational form of the source condition. Each spike is recovered by one or two atoms whose heights match the amplitude of the ground truth up to the standard $\sim 10\text{--}25\%$ $\mathcal{O}(\lambda)$ BLASSO shrinkage (positions exact to $\sim 10^{-3}$).

Theorem 1 applied at $(\mu_k, \tilde{h}_k)$ then delivers the finite, computable Fenchel–Young duality-gap decomposition

$$\Delta_k = \underbrace{L_F(\Phi \mu_k \| \tilde{h}_k)}_{\geq 0, = 0 \iff c_k = 1} + \underbrace{\lambda(\|\mu_k\|_{\text{TV}} - \langle \tilde{\eta}_k, \mu_k \rangle)}_{\geq 0, \text{ finite}} \geq J(\mu_k) - \inf J. \quad (5.2a)$$

The data-fidelity piece L_F in (5.2a) is generically non-zero off Mirror Alignment: trading exact Mirror Alignment ($L_F = 0$) for dual feasibility ($\tilde{h}_k \in K_\lambda$) is what produces a finite bound, and is unavoidable. In the least-squares case, the trade-off is $L_F(\Phi \mu_k \| \tilde{h}_k) = (1 - c_k)^2 F(\Phi \mu_k)$, which vanishes as $c_k \rightarrow 1$, i.e. once the iterate becomes dual-feasible, so $\Delta_k \rightarrow 0$ requires both pieces to vanish jointly.

Figure 3 illustrates the oracle-free guarantee. The red curve Δ_k is computed from the iterate alone, yet it lies above the blue optimality gap $J(\mu_k) - \hat{J}^*$ at every iteration (this is the inequality $\Delta_k \geq J(\mu_k) - \inf J$ of (2.9)), and the two curves merge as $c_k \rightarrow 1$, so the certificate becomes asymptotically tight. Once the data-fidelity piece has vanished ($c_k = 1$, $L_F = 0$), the gap reduces to the single regularizer term λL_R , exactly the quantity whose sub-tolerance level Theorem 4 converts into a feasible Brøndsted–Rockafellar proxy. A practitioner targeting accuracy ϵ may therefore stop as soon as $\Delta_k \leq \epsilon$ and certify $J(\mu_k) - \inf J \leq \epsilon$ with no oracle. This is the early-stopping rule that the bound is meant to provide, certified by Theorem 1 and (2.9). The reference value $\hat{J}^*$ used to draw the blue curve is itself an oracle quantity (the minimal objective over an extended run); it is shown only to exhibit the bound, and plays no role in the certificate Δ_k , which the optimizer computes online.

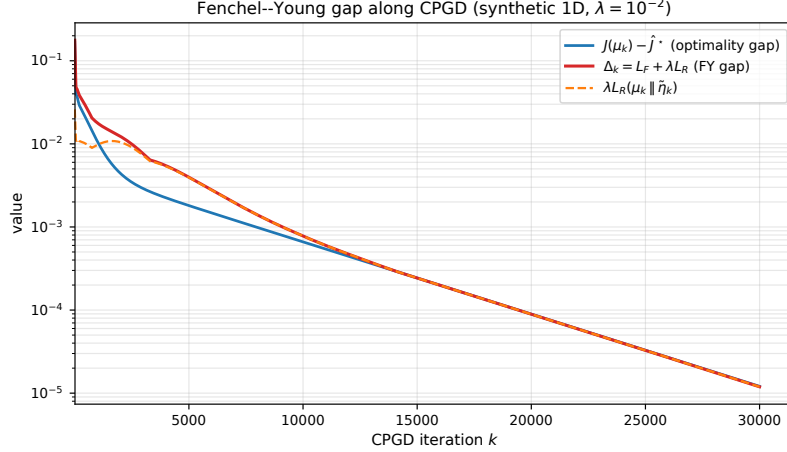


Figure 3: Fenchel–Young duality-gap decomposition $\Delta_k = L_F + \lambda L_R$ (Theorem 1) at the rescaled Mirror-Alignment dual $\hat{h}_k = c_k \nabla F(\Phi \mu_k)$, along the trajectory of Figure 2 ($T = 30\,000$ iterations; semilog axes). Red: total Fenchel–Young gap Δ_k ; orange: the regularizer piece $\lambda L_R(\mu_k \parallel \tilde{\eta}_k)$, which tracks Δ_k once the data-fidelity piece $L_F = (1 - c_k)^2 F(\Phi \mu_k)$ has vanished ($c_k \rightarrow 1$); blue: primal optimality gap $J(\mu_k) - \hat{J}^*$, with $\hat{J}^*$ the minimal objective over an extended reference run (twice the displayed horizon, minimum taken over the tail) as a proxy for $\inf J$.

5.2 Certifiability of Lion-K and Muon via the duality-gap decomposition

The Fenchel–Young duality-gap decomposition $\Delta(\mu, h) = L_F(\Phi \mu \parallel h) + \lambda L_R(\mu \parallel \eta)$ of Theorem 1 is optimizer-agnostic: it holds for any iterate μ_t produced by any optimizer solving the regularized program $\min_{\mu} F(\Phi \mu) + \lambda R(\mu)$, be it CPGD, ISTA, SVT, Lion-K, or Muon. Evaluated at the *rescaled* Mirror-Alignment dual $\hat{h}_t = c_t \nabla F(\Phi \mu_t)$ with $c_t = \min(1, \lambda / \|\Phi^* \nabla F(\Phi \mu_t)\|_{R^*})$, where $\|\cdot\|_{R^*}$ is the norm whose unit ball is $\text{dom } R^*$ (the ℓ_∞ norm for ℓ_1 regularization, the operator norm for the nuclear norm), so that $\hat{h}_t$ is dual-feasible, the resulting Fenchel–Young gap $\Delta_t = L_F(\Phi \mu_t \parallel \hat{h}_t) + \lambda L_R(\mu_t \parallel \tilde{\eta}_t) \geq J(\mu_t) - \inf J$ (2.9) is finite, non-negative, and computable from the iterate alone.

We apply this diagnostic to two non-convex optimizers used in the deep-learning community. Chen et al. [2024] introduced the Lion-K family, whose update applies the subgradient of a convex map K (the sign nonlinearity for Lion); Jordan et al. [2024] introduced Muon, whose update applies the matrix-sign $\text{matsign}(G) = UV^\top$. Both nonlinearities are subgradients of a regularizer: $\text{sign}(x) \in \partial \|x\|_1$ and $UV^\top \in \partial \|U\Sigma V^\top\|_*$. When this sign / matrix-sign step is used as the regularizer’s *proximal* operator (soft-thresholding $\text{soft}_\tau(z) = \text{sign}(z) \max(|z| - \tau, 0)$ for ℓ_1 , and singular-value thresholding $U \text{soft}_\tau(\Sigma) V^\top$ for the nuclear norm, with the polar factor $UV^\top$ supplied by Muon’s matrix-sign / Polar-Express step), the resulting iteration is proximal gradient on $\min_{\mu} F(\Phi \mu) + \lambda R(\mu)$, namely ISTA and SVT, respectively. In this proximal form Lion-K and Muon solve exactly the $F + \lambda R$ program, and the Fenchel–Young gap certifies them. Figure 1 reports this on two ill-conditioned instances (a coherent, AR(1)-correlated, ℓ_1 compressed-sensing problem and a nuclear-norm matrix-sensing problem), chosen so that convergence is gradual rather than instantaneous: for both optimizers the oracle-free Δ_t decays steadily to zero while staying above the (oracle) optimality gap $J(\mu_t) - \hat{J}^*$ along the trajectory, so the certification stays visible over the 10 000 iterations shown.

In summary, Lion-K and Muon share the same algorithmic template, a sign / matrix-sign step, which the source condition of Section 2 (condition (2.10)) identifies as the regularizer subgradient for ℓ_1 and for the nuclear norm. Used as the proximal operator of that regularizer, the step yields proximal gradient on $\min_{\mu} F(\Phi \mu) + \lambda R(\mu)$, and the duality-gap decomposition of Figure 1 certifies, oracle-free, that the gap $\Delta_t \rightarrow 0$, hence that the $F + \lambda R$ program is solved. The companion notebook contains the full Lion-K / Muon / Newton–Schulz numerical implementation; we refer to Amsel et al. [2026] for the polar-expansion analysis underlying the quintic Newton–Schulz iteration that approximates matsign in our Muon implementation.

Remark 12 (Convex certificate for a non-convex deployment). *In their original deep-learning use, Lion-K and Muon are run on non-convex training objectives, whereas the certificate of this section targets the*

convex proximal-gradient surrogate $\min_{\mu} F(\Phi\mu) + \lambda R(\mu)$ that their ISTA/SVT proximal form solves: Δ_t certifies optimality for that convex program, not for the non-convex training loss. This convex program is the informative object rather than a limitation, since by the mean-field/over-parameterization theory of Chizat [2022], Chizat and Bach [2018] over-parameterized non-convex gradient dynamics can realize the global minimizer of a convex problem lifted to the space of measures, the same style of lifting that underlies the CPGD trajectories of Section 5.1, see [De Castro et al., 2023b].

The experiments make the same point in each setting. The oracle-free Fenchel–Young gap $\Delta = L_F + \lambda L_R$ of Theorem 1 upper-bounds the primal optimality gap (2.9) using only the running iterate (Figures 2–3): it tracks the true gap, becomes asymptotically tight as the rescaling contraction $c_k \rightarrow 1$, and supplies the early-stopping rule “stop once $\Delta_k \leq \epsilon$ ”. Because the construction is optimizer-agnostic, the same scalar certifies the deep-learning optimizers Lion-K and Muon: in proximal form their sign / matrix-sign step is the regularizer proximal operator, they solve $F + \lambda R$, and the gap $\Delta_t \rightarrow 0$ along the trajectory (Figure 1). For continuous BLASSO and for finite-dimensional sparse / low-rank recovery, certifying optimality reduces to computing a single scalar.

Conclusion and perspectives

The Banach-space setting admits several relaxations. The measure space $\mathcal{M}(\mathcal{X})$ may be replaced by any dual Banach space $\mathcal{B}^*$. Relaxing the Hilbert space $\mathcal{H}$ to a reflexive Banach space would additionally require a counterpart of Carlier’s inequality, which is where the Hilbert structure enters. A finer quantitative tracking of the Brøndsted–Rockafellar proxy under stochastic optimization noise is left for future work. Building on the error bounds established here, the companion paper [Aubin-Frankowski and De Castro, 2026] develops support localization and exact support recovery for the GBL under a non-degenerate source condition (verified through a Fisher-reweighted Local Positive Curvature pivot, with a Kernel Switch supplying sign-free localization), and attains the parametric statistical rate, up to a logarithmic factor, on the well-specified population logistic GBL.

Disclosure of AI use

In line with the Leiden Declaration on Artificial Intelligence and Mathematics [Leiden Declaration, 2026], we disclose the use of AI agents (Claude Code, Anthropic) in the preparation of this paper. These agents were used to audit the statements and the proofs, through repeated adversarial verification passes recorded in a change-log, and to help with editorial tasks: consistency of notation, bibliography formatting, and typesetting. The authors produced and checked every result and every proof themselves and take sole responsibility for the correctness of the paper and for its citations.

References

- Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://openreview.net/forum?id=yRtgZ1K8h0>.
- Francisco Andrade, Gabriel Peyré, and Clarice Poon. Sparsistency for inverse optimal transport. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/forum?id=wpXGPCBOTX>.
- Francisco Andrade, Gabriel Peyré, and Clarice Poon. Learning from samples: inverse problems over measures. *arXiv preprint arXiv:2505.07124*, 2025. doi: 10.48550/arXiv.2505.07124.
- Hedy Attouch and Haïm Brezis. Duality for the sum of convex functions in general Banach spaces. In Jorge Alberto Barroso, editor, *Aspects of Mathematics and its Applications*, volume 34 of *North-Holland Mathematical Library*, pages 125–133. North-Holland, Amsterdam, 1986. URL <https://sites.math.rutgers.edu/~brezis/PUBlications/106.pdf>.
- Pierre-Cyril Aubin-Frankowski and Yohann De Castro. Fenchel–Young duality gaps: Localization and exact support recovery for the generalized Beurling–Lasso. *arXiv preprint*, 2026. Companion paper.

- Jean-Marc Azaïs, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.
- Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.
- Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.
- Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric de Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.
- Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013. doi: 10.1051/cocv/2011205.
- A. Brondsted and R. T. Rockafellar. On the subdifferentiability of convex functions. *Proceedings of the American Mathematical Society*, 16(4):605, August 1965. ISSN 0002-9939. doi: 10.2307/2033889. URL <http://dx.doi.org/10.2307/2033889>.
- Martin Burger and Stanley Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004. doi: 10.1088/0266-5611/20/5/005.
- Guillaume Carlier. Fenchel–Young inequality with a remainder and applications to convex duality and optimal transport. *SIAM Journal on Optimization*, 33(3):1463–1472, 2023.
- Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As Lyapunov predicts. In *International Conference on Learning Representations (ICLR)*, 2024.
- Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1):487–532, 2022.
- Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Yohann De Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and Applications*, 395(1):336–354, 2012.
- Yohann De Castro, Sébastien Gadat, Clément Marteau, and Cathy Maugis-Rabusseau. SuperMix: sparse regularization for mixtures. *The Annals of Statistics*, 49(3):1779–1809, 2021.
- Yohann De Castro, Vincent Duval, and Romain Petit. Towards off-the-grid algorithms for total variation regularized inverse problems. *Journal of Mathematical Imaging and Vision*, 65(1):53–81, 2023a.
- Yohann De Castro, Sébastien Gadat, and Clément Marteau. FastPart: Over-parameterized stochastic gradient descent for sparse optimisation on measures. *arXiv preprint arXiv:2312.05993*, 2023b.
- Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.
- Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.
- Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces, Volume I: Martingales and Littlewood–Paley Theory*, volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer, Cham, 2016.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, <https://kellerjordan.github.io/posts/muon/>, 2024.

- Leiden Declaration. The Leiden Declaration on Artificial Intelligence and Mathematics. <https://leidendeclaration.ai/>, 2026. DOI: 10.5281/zenodo.20302944. Endorsed by the International Mathematical Union.
- Yves Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, volume 22 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2001.
- Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, 23(1):241–327, 2023. doi: 10.1007/s10208-021-09545-5.
- Walter Rudin. *Real and Complex Analysis*. McGraw–Hill, New York, 3rd edition, 1987.
- Iyad Walwil and Olivier Fercoq. The smoothed duality gap as a stopping criterion. *Mathematical Programming Computation*, 17(4):653–697, May 2025. ISSN 1867-2957. doi: 10.1007/s12532-025-00284-0. URL <http://dx.doi.org/10.1007/s12532-025-00284-0>.
- Constantin Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.

A Technical lemmas

We here regroup technical lemmas underlying some of our proofs. They sometimes recover well-known results in convex analysis, however, since our setting is naturally non-reflexive Banach spaces, it is easier to reprove the formulations we need rather than quote existing statements, often restricted to the Hilbert case in textbooks.

A.1 On the forward operator

Lemma 8. *Let $\mathcal{B}$ be a Banach space and denote $\mathcal{B}^*$ its dual space. Let $\mathcal{H}$ be a Hilbert space. Let $\Phi : \mathcal{B}^* \rightarrow \mathcal{H}$ be a linear weak-*to-weak continuous operator. Then its pre-adjoint $\Phi^* : \mathcal{H} \rightarrow \mathcal{B}$ reads*

$$\Phi^* : h \in \mathcal{H} \mapsto (b^* \in \mathcal{B}^* \mapsto \langle \varphi_{b^*}, h \rangle_{\mathcal{H}}) \in \mathcal{B},$$

where we identified the pre-dual space $\mathcal{B}$ with a subspace of the bidual $\mathcal{B}^{**}$, and where $\varphi_{b^*} := \Phi b^*$ for $b^* \in \mathcal{B}^*$.

Proof. Fix $h \in \mathcal{H}$ and consider the linear functional $\ell_h : \mathcal{B}^* \rightarrow \mathbb{R}$ defined by

$$\ell_h(b^*) := \langle h, \Phi b^* \rangle_{\mathcal{H}}, \quad b^* \in \mathcal{B}^*.$$

It is linear. Since $\Phi : (\mathcal{B}^*, \sigma(\mathcal{B}^*, \mathcal{B})) \rightarrow \mathcal{H}$ is weak-*to-weak continuous by assumption, and $y \mapsto \langle h, y \rangle_{\mathcal{H}}$ is a weakly continuous linear functional on $\mathcal{H}$, the composition ℓ_h is a *weak-* continuous* linear functional on $\mathcal{B}^*$. A standard result in functional analysis states that the dual of $(\mathcal{B}^*, \sigma(\mathcal{B}^*, \mathcal{B}))$ is exactly the canonical identification $\mathcal{B} \hookrightarrow \mathcal{B}^{**}$, hence every weak-* continuous linear functional on $\mathcal{B}^*$ is of the form

$$b^* \mapsto \langle b, b^* \rangle_{\mathcal{B}, \mathcal{B}^*} \quad \text{for a unique } b \in \mathcal{B}.$$

Therefore there exists a unique $\Phi^* h \in \mathcal{B}$ such that, for all $b^* \in \mathcal{B}^*$, $\langle h, \Phi b^* \rangle_{\mathcal{H}} = \langle \Phi^* h, b^* \rangle_{\mathcal{B}, \mathcal{B}^*}$. This defines a linear operator $\Phi^* : \mathcal{H} \rightarrow \mathcal{B}$. Moreover, we obtain the norm estimate

$$\|\Phi^* h\|_{\mathcal{B}} = \sup_{\|b^*\| \leq 1} |\langle \Phi^* h, b^* \rangle| = \sup_{\|b^*\| \leq 1} |\langle h, \Phi b^* \rangle_{\mathcal{H}}| \leq \|h\|_{\mathcal{H}} \|\Phi\|,$$

which shows that Φ^* is bounded. Finally, set $\varphi_{b^*} := \Phi b^*$ for $b^* \in \mathcal{B}^*$. The previous identity can be written

$$\langle \Phi^* h, b^* \rangle_{\mathcal{B}, \mathcal{B}^*} = \langle h, \varphi_{b^*} \rangle_{\mathcal{H}}, \quad \forall b^* \in \mathcal{B}^*,$$

which proves the claimed expression of Φ^* via the canonical identification $\mathcal{B} \hookrightarrow \mathcal{B}^{**}$. $\square$

Remark 13 (Φ^* in the measure case). $\Phi^* h$ is an element of the pre-dual $\mathcal{B}$ of Lemma 8. When $\mathcal{B}^* = \mathcal{M}(\mathcal{X})$, $\mathcal{M}(\mathcal{X})$ is generated (in the weak-* topology) by Dirac measures, so $\Phi^* h$ is fully specified by its values on Dirac inputs: $(\Phi^* h)(t) = \langle \varphi_t, h \rangle_{\mathcal{H}}$ where $\varphi_t := \Phi \delta_t$.

Lemma 9. *Let $\Phi : \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{H}$ be a linear weak-*to-weak continuous operator, with $\mathcal{H}$ a separable Hilbert space. Then its pre-adjoint $\Phi^* : \mathcal{H} \rightarrow \mathcal{C}_0(\mathcal{X})$ reads*

$$\Phi^* : h \in \mathcal{H} \mapsto (t \in \mathcal{X} \mapsto \langle \varphi_t, h \rangle_{\mathcal{H}}) \in \mathcal{C}_0(\mathcal{X}),$$

where we identified the pre-dual space $\mathcal{C}_0(\mathcal{X})$ with a subspace of the dual $\mathcal{M}(\mathcal{X})^*$, and where $\varphi_t := \Phi \delta_t$ with δ_t the Dirac measure at $t \in \mathcal{X}$. Moreover, for every $\mu \in \mathcal{M}(\mathcal{X})$, one has the Bochner integral representation in $\mathcal{H}$:

$$\Phi \mu = \int_{\mathcal{X}} \varphi_t d\mu(t).$$

Proof. Applying Lemma 8 with $\mathcal{B} = \mathcal{C}_0(\mathcal{X})$, $\mathcal{B}^* = \mathcal{M}(\mathcal{X})$ (Riesz–Markov, [Rudin, 1987](#), Theorem 6.19) gives a unique $\Phi^* h \in \mathcal{C}_0(\mathcal{X})$ such that $\langle h, \Phi \mu \rangle_{\mathcal{H}} = \langle \Phi^* h, \mu \rangle_{\mathcal{C}_0(\mathcal{X}), \mathcal{M}(\mathcal{X})}$ for every $\mu \in \mathcal{M}(\mathcal{X})$, with $\|\Phi^* h\|_{\infty} \leq \|h\|_{\mathcal{H}} \|\Phi\|$. Setting $\varphi_t := \Phi \delta_t \in \mathcal{H}$ and evaluating the identity at $\mu = \delta_t$ yields $(\Phi^* h)(t) = \langle \varphi_t, h \rangle_{\mathcal{H}}$, so $\Phi^* h$ is precisely the continuous function $t \mapsto \langle \varphi_t, h \rangle_{\mathcal{H}}$.

We now prove the integral representation. First, note that

$$\|\varphi_t\|_{\mathcal{H}} = \|\Phi \delta_t\|_{\mathcal{H}} \leq \|\Phi\| \|\delta_t\|_{\text{TV}} = \|\Phi\|$$

for all $t \in \mathcal{X}$, so $t \mapsto \varphi_t \in \mathcal{H}$ is bounded. Next, for each fixed $h \in \mathcal{H}$, we already identified $\Phi^*h \in \mathcal{C}_0(\mathcal{X})$ and established $(\Phi^*h)(t) = \langle \varphi_t, h \rangle_{\mathcal{H}}$. Since $\Phi^*h \in \mathcal{C}_0(\mathcal{X})$, the scalar map $t \mapsto (\Phi^*h)(t) = \langle \varphi_t, h \rangle_{\mathcal{H}}$ is continuous on $\mathcal{X}$. Hence $t \mapsto \varphi_t$ is weakly continuous (i.e., all scalar evaluations $\langle \varphi_t, h \rangle_{\mathcal{H}}$ are continuous in t). Because $\mathcal{H}$ is separable, weak measurability/continuity implies strong (Bochner) measurability by Pettis' theorem, see [Hytönen et al., 2016, Theorem 1.1.6]. Consequently, since $\int_{\mathcal{X}} \|\varphi_t\|_{\mathcal{H}} d|\mu|(t) \leq \|\Phi\| \|\mu\|_{\text{TV}} < \infty$, for any finite signed measure $\mu \in \mathcal{M}(\mathcal{X})$ the Bochner integral $\int_{\mathcal{X}} \varphi_t d\mu(t)$ is well-defined in $\mathcal{H}$ and satisfies, for all $h \in \mathcal{H}$,

$$\left\langle h, \int_{\mathcal{X}} \varphi_t d\mu(t) \right\rangle_{\mathcal{H}} = \int_{\mathcal{X}} \langle h, \varphi_t \rangle_{\mathcal{H}} d\mu(t).$$

Using the identity $\langle h, \Phi\mu \rangle_{\mathcal{H}} = \langle \Phi^*h, \mu \rangle_{\mathcal{C}_0(\mathcal{X}), \mathcal{M}(\mathcal{X})} = \int_{\mathcal{X}} (\Phi^*h)(t) d\mu(t) = \int_{\mathcal{X}} \langle h, \varphi_t \rangle_{\mathcal{H}} d\mu(t)$, we deduce that

$$\langle h, \Phi\mu \rangle_{\mathcal{H}} = \left\langle h, \int_{\mathcal{X}} \varphi_t d\mu(t) \right\rangle_{\mathcal{H}}$$

for all $h \in \mathcal{H}$. By uniqueness of the Riesz representation in $\mathcal{H}$, this implies

$$\Phi\mu = \int_{\mathcal{X}} \varphi_t d\mu(t),$$

which is the desired representation. $\square$

A.2 On the Fenchel–Young losses

Lemma 10 (Properties of Fenchel–Young losses on Banach spaces; cf. Blondel et al., 2020). *Let $\mathcal{B}$ be a Banach space and $\mathcal{B}^*$ its dual space. Let $G : \mathcal{B}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex function on the Banach space $\mathcal{B}^*$ that is weak-* lower semi-continuous ($\sigma(\mathcal{B}^*, \mathcal{B})$ -l.s.c., equivalently $G = G^{**}$ with G^* taken on the pre-dual $\mathcal{B}$). Let G^* be the Legendre–Fenchel conjugate of G given only on the pre-dual space $\mathcal{B}$, identified with a weak-* dense subset of the bidual $\mathcal{B}^{**}$. For any $u \in \mathcal{B}^*$ and $v \in \mathcal{B}$, the following properties hold:*

1. **Non-negativity:** $L_G(u \parallel v) := G(u) + G^*(v) - \langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} \geq 0$.
2. **Zero-loss equivalence:** The following statements are unconditionally equivalent for any $(u, v) \in \mathcal{B}^* \times \mathcal{B}$:
 - (i) $L_G(u \parallel v) = 0$
 - (ii) $v \in \partial G(u)$
 - (iii) $u \in \partial G^*(v)$
3. **Differentiable case:** If G is Gâteaux differentiable at u with $\nabla G(u) \in \mathcal{B}$, then $L_G(u \parallel v) = 0 \iff v = \nabla G(u)$.

Proof. The non-negativity (1) is a direct consequence of the Fenchel–Young inequality, which states that $G^*(v) \geq \langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} - G(u)$ by definition of the supremum in the conjugate.

For the equivalence (2), we proceed as follows:

- (ii) $\implies$ (i): If $v \in \partial G(u)$, then by definition of the subgradient, $G(w) \geq G(u) + \langle v, w - u \rangle_{\mathcal{B}, \mathcal{B}^*}$ for all w . Rearranging implies $\langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} - G(u) \geq \langle v, w \rangle_{\mathcal{B}, \mathcal{B}^*} - G(w)$ for all w . Taking the supremum over w on the RHS gives $\langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} - G(u) \geq G^*(v)$, or equivalently $G(u) + G^*(v) - \langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} \leq 0$. Since the loss is always non-negative, it must be 0.
- (i) $\implies$ (ii): If $G(u) + G^*(v) - \langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} = 0$, then $\langle v, u \rangle_{\mathcal{B}, \mathcal{B}^*} - G(u) = G^*(v)$. By the Fenchel–Young inequality we have $G^*(v) \geq \langle v, w \rangle_{\mathcal{B}, \mathcal{B}^*} - G(w)$ for any w . Combining both leads to the inequality $G(w) \geq G(u) + \langle v, w - u \rangle_{\mathcal{B}, \mathcal{B}^*}$ for any w which proves that $v \in \partial G(u)$.

The equivalence (i) $\Leftrightarrow$ (iii) follows by symmetry from the (i) $\Leftrightarrow$ (ii) argument applied to G^* , using the Fenchel–Moreau identity $G^{**} = G$ on the dual pairing $(\mathcal{B}, \mathcal{B}^*)$, which holds by the weak-* lower semi-continuity of G (no reflexivity of $\mathcal{B}$ needed).

For (3), if G is differentiable, the subdifferential contains a single element, $\partial G(u) = \{\nabla G(u)\}$. Thus condition (ii) simplifies to $v = \nabla G(u)$. See Blondel et al. [2020, Proposition 1] for further details on these convex analysis identities in the context of machine learning losses. $\square$

Lemma 11 (Bregman divergence identity). *Let $F : \mathcal{H} \rightarrow \mathbb{R}$ be a strictly convex, continuously differentiable function. Let $F^* : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ be its Fenchel conjugate, which is differentiable on the interior of its domain. For any $u \in \mathcal{H}$ and any $h \in \text{int}(\text{dom}(F^*))$, the following identity holds:*

$$F(u \mid \nabla F^*(h)) = F^*(h \mid \nabla F(u)) = L_F(u \parallel h).$$

Here $\nabla F^*(\nabla F(u))$ is understood as u , the unique element of $\partial F^*(\nabla F(u))$ (a singleton by strict convexity of F); with this reading the middle term is well defined even when $\nabla F(u) \notin \text{int}(\text{dom } F^*)$.

Proof. 1. Proving $F(u \mid \nabla F^(h)) = L_F(u \parallel h)$:* Let $y = \nabla F^*(h)$. By the properties of Legendre–Fenchel conjugates for strictly convex functions, the gradient maps are inverses: $\nabla F(y) = \nabla F(\nabla F^*(h)) = h$. Substituting $y = \nabla F^*(h)$ and $\nabla F(y) = h$ into the Bregman definition:

$$\begin{aligned} F(u \mid \nabla F^*(h)) &= F(u) - F(\nabla F^*(h)) - \langle h, u - \nabla F^*(h) \rangle \\ &= F(u) - F(\nabla F^*(h)) - \langle h, u \rangle + \langle h, \nabla F^*(h) \rangle. \end{aligned}$$

Recall the Fenchel equality for the pair $(h, \nabla F^*(h))$: since $\nabla F^*(h)$ is the subgradient at h , we have $F^*(h) + F(\nabla F^*(h)) = \langle h, \nabla F^*(h) \rangle$, which implies $F(\nabla F^*(h)) = \langle h, \nabla F^*(h) \rangle - F^*(h)$. Substituting:

$$\begin{aligned} F(u \mid \nabla F^*(h)) &= F(u) - (\langle h, \nabla F^*(h) \rangle - F^*(h)) - \langle h, u \rangle + \langle h, \nabla F^*(h) \rangle \\ &= F(u) + F^*(h) - \langle u, h \rangle = L_F(u \parallel h). \end{aligned}$$

2. Proving $F^*(h \mid \nabla F(u)) = L_F(u \parallel h)$. With the stated convention, $\nabla F^*(\nabla F(u)) = u$: indeed $u \in \partial F^*(\nabla F(u))$ by the Fenchel equality at the pair $(u, \nabla F(u))$, and strict convexity of F makes this subgradient unique. By definition,

$$F^*(h \mid \nabla F(u)) = F^*(h) - F^*(\nabla F(u)) - \langle u, h - \nabla F(u) \rangle.$$

The Fenchel equality at $(u, \nabla F(u))$ gives $F^*(\nabla F(u)) = \langle u, \nabla F(u) \rangle - F(u)$. Substituting yields

$$F^*(h \mid \nabla F(u)) = F^*(h) + F(u) - \langle u, h \rangle = L_F(u \parallel h).$$

Combining the two equalities yields the claimed identity. $\square$

A.3 On the pulled-back regularizer

Lemma 12 (Regularity of the pulled-back regularizer). *Let Assumption 1 hold, and let R^* be the Legendre–Fenchel conjugate of R taken on the pre-dual pairing $(\mathcal{B}, \mathcal{B}^*)$,*

$$R^*(f) = \sup_{w \in \mathcal{B}^*} \{ \langle f, w \rangle_{\mathcal{B}, \mathcal{B}^*} - R(w) \}, \quad f \in \mathcal{B}.$$

Let u be the pulled-back regularizer of (3.1). Then the following hold.

- (a) $R^* : \mathcal{B} \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper, convex and norm lower semi-continuous, and $R^{**} = R$. In particular $R^* > -\infty$ everywhere on $\mathcal{B}$ and $u > -\infty$ everywhere on $\mathcal{H}$.
- (b) u is convex on $\mathcal{H}$.
- (c) u is lower semi-continuous on $\mathcal{H}$.
- (d) The effective domain is

$$\text{dom } u = \left\{ h \in \mathcal{H} : -\frac{\Phi^* h}{\lambda} \in \text{dom } R^* \right\} = (\Phi^*)^{-1}(-\lambda \text{dom } R^*),$$

and u is proper if and only if $\text{dom } u \neq \emptyset$; under Assumption 1 the map u is therefore proper.

If $0 \in \text{dom } R^*$ then $0 \in \text{dom } u$, so the feasibility requirement $\text{dom } u \neq \emptyset$ holds automatically. This is the case for every norm regularizer $R = \|\cdot\|_{\mathcal{B}^*}$, whose conjugate on the pre-dual is $R^* = \iota_{B_{\mathcal{B}}}$, the indicator of the closed unit ball $B_{\mathcal{B}} = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq 1\}$ of the pre-dual, so that $0 \in \text{dom } R^*$. For $R = \|\cdot\|_{\text{TV}}$ on $\mathcal{B}^* = \mathcal{M}(\mathcal{X})$, where $\mathcal{B} = \mathcal{C}_0(\mathcal{X})$, this reads $R^* = \iota_{\{f \in \mathcal{C}_0(\mathcal{X}) : \|f\|_{\infty} \leq 1\}}$ (Section 4). Consequently u is a proper, convex and lower semi-continuous function on the Hilbert space $\mathcal{H}$, which is exactly the regularity required to apply Carlier's constructive Brøndsted–Rockafellar theorem [Carlier, 2023, Theorem 4.1] in Theorem 4.

We do not use, and do not have, coercivity of u . For a norm regularizer $u = \lambda \iota_{B_{\mathcal{B}}} \circ A$ is the indicator of the set $K_{\lambda} = \{h \in \mathcal{H} : \|\Phi^* h\|_{\mathcal{B}} \leq \lambda\}$, which contains $\text{Ker } \Phi^*$ and is unbounded whenever Φ^* is not bounded below, so u does not tend to $+\infty$ as $\|h\|_{\mathcal{H}} \rightarrow \infty$. What the downstream argument needs is only that u is proper, convex and lower semi-continuous, equivalently that $u^{**} = u$ and u^* is proper, which is what the value $u^*(-\Phi\mu)$ in Proposition 2 uses.

Proof. Throughout, the pairing $\langle f, w \rangle_{\mathcal{B}, \mathcal{B}^*}$ is between $f \in \mathcal{B}$ and $w \in \mathcal{B}^*$, and each $w \in \mathcal{B}^*$ is a bounded linear functional on $\mathcal{B}$ with $|\langle f, w \rangle_{\mathcal{B}, \mathcal{B}^*}| \leq \|f\|_{\mathcal{B}} \|w\|_{\mathcal{B}^*}$. Write $A : \mathcal{H} \rightarrow \mathcal{B}$ for the linear map:

$$Ah := -\Phi^* h / \lambda, \text{ so that } u = \lambda (R^* \circ A).$$

Proof of (a). For each $w \in \text{dom } R$ the map $f \mapsto \langle f, w \rangle_{\mathcal{B}, \mathcal{B}^*} - R(w)$ is affine and norm continuous on $\mathcal{B}$, since $f \mapsto \langle f, w \rangle_{\mathcal{B}, \mathcal{B}^*}$ is a bounded linear functional and $-R(w)$ is a finite constant. By definition R^* is the pointwise supremum of this family. A pointwise supremum of continuous affine functionals is convex and norm lower semi-continuous, so R^* is convex and norm lower semi-continuous on $\mathcal{B}$, and its epigraph $\text{epi } R^* = \bigcap_{w \in \text{dom } R} \text{epi}[\langle \cdot, w \rangle_{\mathcal{B}, \mathcal{B}^*} - R(w)]$ is a closed convex subset of $\mathcal{B} \times \mathbb{R}$. Now, fix $w_0 \in \text{dom } R$. Then for every $f \in \mathcal{B}$,

$$R^*(f) \geq \langle f, w_0 \rangle_{\mathcal{B}, \mathcal{B}^*} - R(w_0) > -\infty,$$

so $R^* > -\infty$ everywhere. It remains to show $R^* \neq +\infty$, and this is the only step that uses the weak-* lower semi-continuity of R , rather than only its convexity. The space $(\mathcal{B}^*, \sigma(\mathcal{B}^*, \mathcal{B}))$ is a locally convex Hausdorff space whose topological dual is exactly $\mathcal{B}$, so the conjugation on the pairing $(\mathcal{B}, \mathcal{B}^*)$ and the biconjugation on $\mathcal{B}^*$ are the ones associated with this duality. As R is proper, convex and $\sigma(\mathcal{B}^*, \mathcal{B})$ -lower semi-continuous, the Fenchel–Moreau theorem [Zălinescu, 2002, Theorem 2.3.3] gives $R^{**} = R$. Hence $\text{dom } R^* \neq \emptyset$, so R^* admits a $\sigma(\mathcal{B}^*, \mathcal{B})$ -continuous affine minorant indexed by $\mathcal{B}$ and is finite somewhere. Together with $R^* > -\infty$, this makes R^* proper. Since $R^* > -\infty$ and $\lambda > 0$, the composition $u = \lambda R^* \circ A$ satisfies $u > -\infty$ everywhere. This proves (a).

Furthermore, by Lemma 8, $\Phi^* : \mathcal{H} \rightarrow \mathcal{B}$ is bounded linear, with $\|\Phi^* h\|_{\mathcal{B}} \leq \|\Phi\| \|h\|_{\mathcal{H}}$. Since $\lambda > 0$, the map $A = -\Phi^* / \lambda$ is linear and norm-to-norm continuous from $\mathcal{H}$ to $\mathcal{B}$, with $\|Ah\|_{\mathcal{B}} \leq (\|\Phi\| / \lambda) \|h\|_{\mathcal{H}}$.

Proof of (b). Convexity of u is the standard fact that a convex function composed with an affine map is convex.

Proof of (c). A function is lower semi-continuous if and only if its epigraph is closed. Define the linear map

$$\Lambda : \mathcal{H} \times \mathbb{R} \rightarrow \mathcal{B} \times \mathbb{R}, \quad \Lambda(h, t) := (Ah, t/\lambda),$$

which is continuous, being the product of the continuous map A from above and the continuous scaling $t \mapsto t/\lambda$. Since $\lambda > 0$, for $(h, t) \in \mathcal{H} \times \mathbb{R}$ we have

$$(h, t) \in \text{epi } u \iff \lambda R^*(Ah) \leq t \iff R^*(Ah) \leq t/\lambda \iff (Ah, t/\lambda) \in \text{epi } R^*,$$

so $\text{epi } u = \Lambda^{-1}(\text{epi } R^*)$. From (a), the set $\text{epi } R^*$ is closed in $\mathcal{B} \times \mathbb{R}$, and Λ is continuous, hence $\text{epi } u$ is closed in $\mathcal{H} \times \mathbb{R}$ and u is lower semi-continuous. Equivalently, each sublevel set $\{u \leq \alpha\} = A^{-1}(\{R^* \leq \alpha/\lambda\})$ is the preimage of a closed convex set under the continuous A , hence closed. The only inputs specific to our setting are the norm continuity of A and the closedness of $\text{epi } R^*$ above.

Proof of (d). Since $R^* > -\infty$ and $\lambda > 0$, the value $u(h) = \lambda R^*(Ah)$ is finite if and only if $R^*(Ah) < +\infty$, that is if and only if $Ah = -\Phi^* h / \lambda \in \text{dom } R^*$. Hence

$$\text{dom } u = \{h \in \mathcal{H} : -\frac{\Phi^* h}{\lambda} \in \text{dom } R^*\} = A^{-1}(\text{dom } R^*) = (\Phi^*)^{-1}(-\lambda \text{dom } R^*),$$

matching (3.1). As u never takes the value $-\infty$ (by (a)), u is proper if and only if it is finite somewhere, that is if and only if $\text{dom } u \neq \emptyset$. Under Assumption 1 we assume $\text{dom } u \neq \emptyset$, so u is proper. $\square$

Lemma 13 (Dual certificates in the range and its closure). *Let Assumption 1 hold, fix $\mu \in \text{dom } R$, and write*

$$s(\mu) := \sup_{f \in \text{ran } \Phi^*} [\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R^*(f)],$$

so that, by Proposition 2, equality holds in (3.2) if and only if $s(\mu) = R(\mu)$. Moreover:

- (i) *If μ admits a dual certificate in the range, i.e. some $f \in \text{ran } \Phi^*$ with $f \in \partial R(\mu)$, then $s(\mu) = R(\mu)$ and equality holds in (3.2).*

Assume in addition that $R = \|\cdot\|_{\mathcal{B}^}$ is a norm regularizer, so that $R^* = \iota_{B_{\mathcal{B}}}$ is the indicator of the closed unit ball $B_{\mathcal{B}} = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq 1\}$ of the pre-dual (Lemma 12(e)) and $s(\mu) = \sup_{f \in \text{ran } \Phi^*, \|f\|_{\mathcal{B}} \leq 1} \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}$; the canonical instance being $R = \|\cdot\|_{\text{TV}}$ on $\mathcal{B}^* = \mathcal{M}(\mathcal{X})$, with $B_{\mathcal{B}} = \{f \in C_0(\mathcal{X}) : \|f\|_{\infty} \leq 1\}$. Then:*

- (ii) *If μ admits a certificate only in the closure, that is $f^* \in \overline{\text{ran } \Phi^*} \cap \partial R(\mu)$ with the closure taken in $\|\cdot\|_{\mathcal{B}}$, then $s(\mu) = R(\mu)$ by rescaling, so equality holds in (3.2). For $R = \|\cdot\|_{\text{TV}}$ this hypothesis is an f^* in the sup-norm closure of $\text{ran } \Phi^*$ with $\|f^*\|_{\infty} \leq 1$ and $\langle f^*, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = \|\mu\|_{\text{TV}}$.*
- (iii) *If $\text{ran } \Phi^*$ is dense in $\mathcal{B}$, then $s(\mu) = R(\mu)$ for every $\mu \in \text{dom } R$, so equality holds in (3.2) throughout $\text{dom } R$.*

Remark 14. *A certificate in $\text{ran } \Phi^*$, or even in $\mathcal{B}$, need not exist, so the converse of (i) is false. For a norm regularizer the subdifferential chain rule $\partial u(\bar{h}) = -\Phi \partial R^*(\bar{\eta})$ nonetheless holds with equality (Proposition 3); the failure here is of certificate existence in $\text{ran } \Phi^*$, not of that equality.*

To see that a certificate need not exist, take $\mathcal{X} = [0, 1]$, $\mathcal{B} = \mathcal{C}(\mathcal{X})$ and $R = \|\cdot\|_{\text{TV}}$, and choose two disjoint sequences $\{a_k\}_{k \geq 1}$, $\{b_k\}_{k \geq 1}$ in $[0, 1]$, each dense in $[0, 1]$. Now, the finite signed measure $\mu := \sum_{k \geq 1} 2^{-k}(\delta_{a_k} - \delta_{b_k})$ has $\|\mu\|_{\text{TV}} = \sum_{k \geq 1} 2^{-k} \cdot 2 = 2$. Any $f \in \partial R(\mu)$ has $\|f\|_{\infty} \leq 1$ and $\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^} = \sum_{k \geq 1} 2^{-k}(f(a_k) - f(b_k)) = 2$; since each summand is at most $2^{-k} \cdot 2$, this forces $f(a_k) = 1$ and $f(b_k) = -1$ for every k . Continuity of f with $f = 1$ on the dense set $\{a_k\}$ gives $f \equiv 1$, while $f = -1$ on the dense set $\{b_k\}$ gives $f \equiv -1$, a contradiction. Hence $\partial R(\mu) \cap \mathcal{B} = \emptyset$: no certificate exists in $\mathcal{B}$, and a fortiori none in $\text{ran } \Phi^*$, even though $s(\mu) = \|\mu\|_{\text{TV}}$ as soon as $\text{ran } \Phi^*$ is dense in $C_0(\mathcal{X})$, by part (iii).*

Proof of Lemma 13. For any $f \in \mathcal{B}$ the Fenchel–Young inequality in the pairing $(\mathcal{B}, \mathcal{B}^*)$ reads $R(\mu) + R^*(f) \geq \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}$, with equality if and only if $f \in \partial R(\mu)$; equivalently

$$f \in \partial R(\mu) \iff \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R^*(f) = R(\mu). \quad (\text{A.1})$$

The bound $s(\mu) \leq R(\mu)$ always holds: it is the inequality chain in the proof of Proposition 2, where the supremum over the subspace $\text{ran } \Phi^*$ is bounded by the one over $\mathcal{B}$, equal to $R^{**}(\mu) = R(\mu)$ (Lemma 12(a)). So in each part it is enough to prove $s(\mu) \geq R(\mu)$.

Part (i). If $f \in \text{ran } \Phi^* \cap \partial R(\mu)$, then by (A.1) the value $\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} - R^*(f) = R(\mu)$ is attained by an f admissible in the supremum defining $s(\mu)$, so $s(\mu) \geq R(\mu)$. This part uses only Assumption 1, and Proposition 2 already contains it; it is recorded here for reference.

For parts (ii) and (iii) let $R = \|\cdot\|_{\mathcal{B}^*}$, so $R^* = \iota_{B_{\mathcal{B}}}$ (Lemma 12(e)). Then (A.1) reduces to $\|f\|_{\mathcal{B}} \leq 1$ and $\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = R(\mu)$, and $s(\mu) = \sup_{f \in \text{ran } \Phi^*, \|f\|_{\mathcal{B}} \leq 1} \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}$, with $\langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} \leq \|f\|_{\mathcal{B}} \|\mu\|_{\mathcal{B}^*} \leq R(\mu)$ on the feasible set.

Part (ii): the rescaling argument. Let $f^* \in \overline{\text{ran } \Phi^*} \cap \partial R(\mu)$, so $\|f^*\|_{\mathcal{B}} \leq 1$ and $\langle f^*, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = R(\mu)$. Fix $\epsilon \in (0, 1)$. As f^* lies in the $\|\cdot\|_{\mathcal{B}}$ -closure of $\text{ran } \Phi^*$, there is $g \in \text{ran } \Phi^*$ with $\|g - f^*\|_{\mathcal{B}} \leq \epsilon$, whence $\|g\|_{\mathcal{B}} \leq \|f^*\|_{\mathcal{B}} + \epsilon \leq 1 + \epsilon$ and

$$\langle g, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = \langle f^*, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} + \langle g - f^*, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} \geq R(\mu) - \|g - f^*\|_{\mathcal{B}} \|\mu\|_{\mathcal{B}^*} \geq (1 - \epsilon)R(\mu).$$

Because $\text{ran } \Phi^*$ is a linear subspace, $\tilde{g} := g/(1 + \epsilon)$ again lies in $\text{ran } \Phi^*$, with $\|\tilde{g}\|_{\mathcal{B}} \leq 1$, so $\tilde{g}$ is admissible for $s(\mu)$ and

$$\langle \tilde{g}, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} = \frac{\langle g, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}}{1 + \epsilon} \geq \frac{1 - \epsilon}{1 + \epsilon} R(\mu).$$

Thus $s(\mu) \geq \frac{1 - \epsilon}{1 + \epsilon} R(\mu)$ for every $\epsilon \in (0, 1)$, and letting $\epsilon \downarrow 0$ gives $s(\mu) \geq R(\mu)$.

Part (iii). Assume $\text{ran } \Phi^*$ is dense in $\mathcal{B}$ and fix $\mu \in \text{dom } R$. By the definition of the dual norm, $R(\mu) = \|\mu\|_{\mathcal{B}^*} = \sup_{\|f\|_{\mathcal{B}} \leq 1} \langle f, \mu \rangle_{\mathcal{B}, \mathcal{B}^*}$, so for $\epsilon \in (0, 1)$ there is $f_\epsilon \in \mathcal{B}$ with $\|f_\epsilon\|_{\mathcal{B}} \leq 1$ and $\langle f_\epsilon, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} \geq (1 - \epsilon)R(\mu)$. Running the argument of Part (ii) with f_ϵ in place of f^* (pick $g \in \text{ran } \Phi^*$ with $\|g - f_\epsilon\|_{\mathcal{B}} \leq \epsilon$ and set $\tilde{g} = g/(1 + \epsilon)$) yields an admissible $\tilde{g}$ with $\langle \tilde{g}, \mu \rangle_{\mathcal{B}, \mathcal{B}^*} \geq \frac{1-2\epsilon}{1+\epsilon}R(\mu)$, so $s(\mu) \geq \frac{1-2\epsilon}{1+\epsilon}R(\mu)$; letting $\epsilon \downarrow 0$ gives $s(\mu) = R(\mu)$ for every $\mu \in \text{dom } R$. $\square$

B Some data-fidelity terms and Fenchel–Young losses

B.1 Least squares (Gaussian noise)

Let $y \in \mathcal{H}$ be the observed data. We consider the standard quadratic loss corresponding to additive Gaussian noise. This loss is finite-valued, $\mathcal{C}^1$, strictly convex and coercive, so it lies in $\mathcal{F}_0$.

Primal: $F(h) = \frac{1}{2}\|h - y\|_{\mathcal{H}}^2$; **Gradient:** $\nabla F(h) = h - y$; **Dual:** $F^*(w) = \frac{1}{2}\|w\|_{\mathcal{H}}^2 + \langle w, y \rangle_{\mathcal{H}}$; **Dual gradient:** $\nabla F^*(w) = w + y$; **Fenchel–Young loss:**

$$L_F(h \parallel w) = \frac{1}{2}\|h - \nabla F^*(w)\|_{\mathcal{H}}^2 = \frac{1}{2}\|h - (w + y)\|_{\mathcal{H}}^2.$$

Minimizing the Fenchel–Young loss over the dual variable w thus targets the residual $h - y$.

B.2 Logistic loss (binary classification)

Let $y \in \{-1, 1\}^n$ be binary labels and $h \in \mathbb{R}^n$ be the logits.

Primal: $F(h) = \sum_{i=1}^n \log(1 + \exp(-y_i h_i))$; **Gradient:** $\nabla F(h)_i = -y_i \sigma(-y_i h_i)$, where $\sigma(t) = (1 + e^{-t})^{-1}$ is the sigmoid function; **Dual:** The conjugate corresponds to the *negative* binary entropy (the convex object a conjugate must be; the binary entropy itself is concave). For $w \in \mathbb{R}^n$ such that $-y_i w_i \in [0, 1]$ for all i , with the convention $0 \log 0 = 0$ (so the boundary values are finite, e.g. $F^*(0) = 0$):

$$F^*(w) = \sum_{i=1}^n [(-y_i w_i) \log(-y_i w_i) + (1 + y_i w_i) \log(1 + y_i w_i)], \quad \text{otherwise, } F^*(w) = +\infty;$$

Dual gradient: $\nabla F^*(w)_i = -y_i \log\left(\frac{-y_i w_i}{1 + y_i w_i}\right)$ (well defined for $-y_i w_i \in (0, 1)$); **Fenchel–Young loss:**

$$L_F(h \parallel w) = \sum_{i=1}^n (\log(1 + e^{-y_i h_i}) + (-y_i w_i) \log(-y_i w_i) + (1 + y_i w_i) \log(1 + y_i w_i) - w_i h_i).$$

This loss F is finite-valued, $\mathcal{C}^1$ and strictly convex but not coercive, since $F \rightarrow 0$ along correctly classified directions, so it lies outside $\mathcal{F}_0$; because $\text{dom } F = \mathcal{H}$, the Fenchel–Rockafellar qualification still gives strong duality and dual attainment. The Fenchel–Young error bounds of Theorem 3 hold as stated, since they need only a proper, l.s.c. convex F and a source-condition certificate (here $\text{dom } F = \mathcal{H}$, so $J_{\mathcal{B}^*}(\mu^*) < +\infty$ is automatic); leaving $\mathcal{F}_0$ costs only the weak-coercivity bound (2.2). Since F is $\mathcal{C}^1$ and strictly convex, the Mirror–Alignment characterization and the uniqueness of the aligned certificate still hold.

B.3 Poisson loss (exponential link)

Let $y \in \mathbb{R}_+^n$ be count data.

Primal: $F(h) = \sum_{i=1}^n (\exp(h_i) - y_i h_i)$; **Gradient:** $\nabla F(h)_i = \exp(h_i) - y_i$; **Dual:** Let $w \in \mathbb{R}^n$. The domain is restricted to $\{w \mid w + y \geq 0\}$ (element-wise),

$$F^*(w) = \sum_{i=1}^n [(w_i + y_i) \log(w_i + y_i) - (w_i + y_i)];$$

Dual gradient: $\nabla F^*(w)_i = \log(w_i + y_i)$; **Fenchel–Young loss:**

$$L_F(h \parallel w) = \sum_{i=1}^n \left(e^{h_i} + (w_i + y_i) \log\left(\frac{w_i + y_i}{e^{h_i}}\right) - (w_i + y_i) \right).$$

Equivalently, $L_F(h \parallel w) = \sum_i D_{\text{KL}}(w_i + y_i \parallel e^{h_i})$ with the dual-shifted observation $w + y$ as the first argument and the predicted intensity e^h as the second; the order matters because D_{KL} is asymmetric.

This formulation is standard for Poisson regression where the intensity parameter is modeled as $\nu = \exp(h)$ to enforce positivity (we write ν for the intensity to avoid collision with the regularization parameter λ). This loss F is finite-valued, $\mathcal{C}^1$ and strictly convex, and it lies in $\mathcal{F}_0$ provided every count is strictly positive ($y > 0$); if some $y_i = 0$ the coercivity bound (2.2) fails as $h_i \rightarrow -\infty$, so F leaves $\mathcal{F}_0$; the loss remains $\mathcal{C}^1$ and strictly convex, so the Mirror-Alignment characterization and the uniqueness of the aligned certificate are unaffected, while the Fenchel-Young error bounds of Theorem 3 still hold given a source-condition certificate ($\text{dom } F = \mathcal{H}$, so $J_{\mathcal{B}^*}(\mu^*) < +\infty$).

B.4 Kullback–Leibler divergence (linear link)

Here we assume the operator Φ maps to a positive domain, $h \in \mathbb{R}_+^n$, representing density or intensity directly.

Primal: $F(h) = \sum_{i=1}^n \left(h_i \log \left(\frac{h_i}{y_i} \right) - h_i + y_i \right)$; **Gradient:** $\nabla F(h)_i = \log(h_i) - \log(y_i)$; **Dual:** $F^*(w) = \sum_{i=1}^n y_i (e^{w_i} - 1)$; **Dual gradient:** $\nabla F^*(w)_i = y_i e^{w_i}$; **Fenchel–Young loss:**

$$L_F(h \parallel w) = \sum_{i=1}^n \left(h_i \log \left(\frac{h_i}{y_i e^{w_i}} \right) - h_i + y_i e^{w_i} \right) = D_{\text{KL}}(h \parallel y e^w).$$

This loss F is the Csiszár I-divergence. Because $\text{dom } F = \mathbb{R}_+^n \subsetneq \mathcal{H}$, this fidelity lies outside $\mathcal{F}_0$. The Fenchel–Young error bounds of Theorem 3 hold given a source-condition certificate, provided $\Phi \mu^* \in \text{dom } F$, that is $J_{\mathcal{B}^*}(\mu^*) < +\infty$: on the positive cone $\{\Phi \mu > 0\}$ the loss F is finite, $\mathcal{C}^1$ and strictly convex, so the Mirror-Alignment and uniqueness statements also apply there.

B.5 Huber loss (robust regression)

The Huber loss is standard in robust statistics for data with outliers. Let $y \in \mathbb{R}^n$ be the observed data. It behaves like a squared error for small residuals and like an absolute error for large residuals. Let $\delta > 0$ be the robustness threshold.

Primal: The function is defined component-wise as $F(h) = \sum_{i=1}^n \rho_\delta(h_i - y_i)$, where:

$$\rho_\delta(r) = \begin{cases} \frac{1}{2} r^2 & \text{if } |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta) & \text{otherwise;} \end{cases}$$

Gradient: The gradient acts as a “soft” saturation function (censoring large residuals):

$$[\nabla F(h)]_i = \begin{cases} h_i - y_i & \text{if } |h_i - y_i| \leq \delta, \\ \delta \cdot \text{sign}(h_i - y_i) & \text{otherwise;} \end{cases}$$

Dual: The convex conjugate is the squared norm restricted to a box constraint, so robust regression is dual to constrained least squares:

$$F^*(w) = \begin{cases} \frac{1}{2} \|w\|_{\mathcal{H}}^2 + \langle w, y \rangle_{\mathcal{H}} & \text{if } \|w\|_{\infty} \leq \delta, \\ +\infty & \text{otherwise;} \end{cases}$$

Dual subdifferential: For interior w (i.e., $\|w\|_{\infty} < \delta$), $\nabla F^*(w)$ is single-valued with $[\nabla F^*(w)]_i = w_i + y_i$. For boundary w with some $|w_i| = \delta$, $\partial F^*(w)$ is set-valued: $\partial F^*(w) = y + w + N_{B_{\infty}^\delta(0)}(w)$, and any selection in this set serves as a Fenchel–Young dual gradient; **Fenchel–Young loss:** The loss $L_F(h \parallel w)$ penalizes the discrepancy between the residual and the saturated dual variable. For any w such that $\|w\|_{\infty} \leq \delta$:

$$L_F(h \parallel w) = \sum_{i=1}^n \left(\rho_\delta(h_i - y_i) + \frac{1}{2} w_i^2 + w_i y_i - w_i h_i \right).$$

If we choose the optimal dual $w = \nabla F(h)$ (the saturated residual), the loss vanishes.

This loss F is $\mathcal{C}^1$ and coercive but not strictly convex, since it is affine in the large-residual regime, so it lies outside $\mathcal{F}_0$. The gap decomposition, strong duality and the Fenchel–Young error bounds of Theorem 3 still hold ($\text{dom } F = \mathcal{H}$); only the optimal prediction $\Phi\mu^\star$ and the Mirror-Alignment identity (2.11) are pinned down up to the flat directions of F .

B.6 Squared hinge loss (L2-SVM)

This loss function is a differentiable alternative to the standard hinge loss, often used to enforce classification margins while maintaining a smooth objective for optimization. Let $y \in \{-1, 1\}^n$ be binary labels.

Primal: $F(h) = \frac{1}{2} \sum_{i=1}^n \max(0, 1 - y_i h_i)^2$; **Gradient:** The gradient is:

$$[\nabla F(h)]_i = -y_i \max(0, 1 - y_i h_i);$$

Dual: The convex conjugate involves a quadratic term restricted to a specific sign constraint. For the dual variable $w \in \mathbb{R}^n$, the domain is restricted to $K_y := \{w \mid -y_i w_i \geq 0 \text{ for all } i\}$ (i.e., w and y must have opposite signs),

$$F^\star(w) = \begin{cases} \sum_{i=1}^n \left(\frac{1}{2} w_i^2 + y_i w_i \right) & \text{if } -y_i w_i \geq 0 \text{ for all } i, \\ +\infty & \text{otherwise;} \end{cases}$$

Dual subdifferential: For w in the relative interior of K_y (i.e., $-y_i w_i > 0$ for every i), $\nabla F^\star$ is single-valued with $[\nabla F^\star(w)]_i = w_i + y_i$. At boundary points where some $w_i = 0$, $\partial F^\star(w) = y + w + N_{K_y}(w)$ is set-valued; the formula below selects one element of this set. **Fenchel–Young loss:** The loss $L_F(h \parallel w)$ measures the gap between the margin error and the dual certificate. For feasible w :

$$L_F(h \parallel w) = \sum_{i=1}^n \left(\frac{1}{2} \max(0, 1 - y_i h_i)^2 + \frac{1}{2} w_i^2 + y_i w_i - w_i h_i \right).$$

It measures how well the dual variable w certifies the classification margin of the prediction h .

This loss F is $\mathcal{C}^1$ but neither coercive nor strictly convex, since it is flat along correctly classified directions, so it lies outside $\mathcal{F}_0$. The Fenchel–Young error bounds of Theorem 3 still hold ($\text{dom } F = \mathcal{H}$), as they do not rely on the aligned base point; only the optimal prediction $\Phi\mu^\star$ and the Mirror-Alignment identity (2.11) are pinned down up to those flat directions.