Consider the Lasso Problem

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1$$

for $\beta \in \mathbb{R}^p$, X as an $n \times p$ matrix. This is a normalized matrix X, usually $\|X\delta_i\| \leq 1$. For classical ML setups, we might have loss functions corresp. to regression or classification. E.g.,

$$\min_{\beta} \frac{1}{n}\sum_i \phi_i(x_i^T\beta) + \lambda\|\beta\|_1$$

where $x_i = \delta_i^T X$. Convex conjugates are:

	$\phi(z)$	$\phi^*(z)$
Squared	$\frac{1}{2}(z-y_i)^2$	$\frac{1}{2}(z+y_i)^2 - \frac{1}{2}y_i^2$
Logistic	$\log(1+\exp(-y_i z))$	$\frac{-z}{y_i}\log\left(\frac{-z}{y_i}\right) + \left(1+\frac{z}{y_i}\right)\log\left(1+\frac{z}{y_i}\right)$

See Blitz: A Principled Meta Algorithm Appendix E for derivation

But the original problem is unconstrained.
So the dual of $\min_\beta L(\beta) + \lambda\|\beta\|_1$
is just $\max_\emptyset p^*$, which is trivial and unhelpful.

So we need to add constraints which will expose some perturbations, such that duals provide a useful optimization view. There will be $\underline{2}$ distinct duals (I tried messing with them, but they seem to be distinct). All of them will have Strong Duality by relaxed Slater

<u>n-parameter dual.</u>

$$\min_{\substack{\beta \in \mathbb{R}^p \\ r \in \mathbb{R}^n}} \frac{1}{2}\|r\|_2^2 + \lambda\|\beta\|_1$$

$$\text{st } X\beta - y = r$$

introduce the dual variable $\nu \in \mathbb{R}^n$

$$\max_{\nu} \inf_{\beta, r} \frac{1}{2}\|r\|_2^2 + \lambda\|\beta\|_1 + \nu^T(X\beta - y - r)$$

$$\to \inf_{\beta} \lambda\|\beta\|_1 + \nu^T X\beta = \begin{cases} 0 & \|X^T\nu\|_\infty \le \lambda \\ -\infty & o/w \end{cases}$$

by looking at $-\beta$ & the convex conjugate of $\|\cdot\|_1$

$$= \max_{\nu} -\nu^T y - 1/2\|\nu\|_2^2 + \inf_{r} \underbrace{\frac{1}{2}\|r\|_2^2 - \nu^T r + \frac{1}{2}\|\nu\|_2^2}_{\underbrace{\frac{1}{2}\|\nu - r\|_2^2}_{0 \text{ @ } r=\nu}}$$

$$\text{st } \|X^T\nu\|_\infty \le \lambda$$

$$= \min_{\nu} -\nu^T y - \frac{1}{2}\|\nu\|_2^2$$

$$\text{st } \|X^T\nu\|_\infty \le \lambda$$

Now lets generalize this for GLMs.

$$\min_{\beta} \frac{1}{n}\sum_i \phi_i(x_i^T\beta) + \lambda\|\beta\|_1$$

Similarly, create constraints:

$$\min_{\beta, z} \frac{1}{n}\sum_i \phi_i(z_i) + \lambda\|\beta\|_1$$

$$\text{st} \qquad X\beta = z$$

$$= \max_{\nu} \min_{\beta, z} \underbrace{\nu^T X\beta + \lambda\|\beta\|_1}_{\text{as before}} + \underbrace{\frac{1}{n}\sum_i \phi_i(z_i) - \nu_i z_i}_{\text{negative of cvx conj.}}$$

$$= \max_{\nu} \frac{1}{n}\sum_i -\phi_i^*(\nu_i) \quad \text{s.t.} \quad \|X^T\nu\|_\infty \le 1$$

$$= -\min_{\nu} \frac{1}{n}\sum_i \phi_i^*(\nu_i)$$

$$\text{st} \quad \|X^T\nu\|_\infty \le 1$$

P-parameter dual

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1$$

$$= \min_{\beta,\gamma} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\gamma\|_1$$

$$\text{s.t. } \gamma = \beta \qquad (\gamma \in \mathbb{R}^p)$$

$$= \max_{\omega} \min_{\beta,\gamma} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\gamma\|_1 + \omega^T(\beta - \gamma)$$

$$= \max_{\omega:\|\omega\|_\infty \le \lambda} \min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \omega^T\beta$$

At this point, let's pause for an interlude.

Note the Moore-Penrose Pseudoinverse X^+.

If the SVD is $U\Sigma V^T = X$, then

$X^+ = V\Sigma^{-1}U^T$, where the SVD here is reduced i.e., Σ is $(\text{rank} X) \times (\text{rank} X)$.

Note the least-norm least squares solution is then $X^+ y = \beta_{LS}$ for minimizing $\|X\beta - y\|_2^2$.

This does not require X to be full rank.

A central property of X^+ is projection:

$X^+ X$ is an orthogonal projection onto $\text{span}(X^T)$ from $\mathbb{R}^p$

This implies $I - X^+ X = \text{proj}_{\ker X}$ so

$X(I - X^+ X) = 0$. Thus we can write any $\beta^* = X^+ \alpha + (I - X^+ X)\delta$ for some α, δ conformably defined.

Notice that then we can write our old optimization in terms of $\alpha \in \mathbb{R}^n$, $\delta \in \mathbb{R}^p$

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + w^T\beta$$

$$= \min_{\alpha, \delta} \frac{1}{2}\|XX^+\alpha - y\|_2^2 + w^T\left(X^+\alpha + (I - X^+X)\delta\right)$$

if $w \notin \text{span}(X^T)$ then $\text{rank}\, X < p$

and $I - X^+X \neq 0$. Choosing

$$\delta = n\left(-(I - X^+X)\right)w$$

results in the obj $\to -\infty$ as $n \to \infty$.

thus this minimum creates an implicit constraint

on w.

This leaves $\min_{\alpha} \|XX^+\alpha - y\|_2^2 + w^T(X^+\alpha)$

as $w \in \text{span}(X^T) \Rightarrow w^T(I - X^+X) = 0$

$$\min_{\alpha} \frac{1}{2}\|XX^+\alpha\|^2 - y^T(XX^+\alpha) + \frac{1}{2}\|y\|_2^2$$
$$+ w^T(X^+\alpha).$$

By projection $X^+ = X^+XX^+$, so

$$= \min_{\alpha} \frac{1}{2}\|XX^+\alpha\|_2^2 - (y-(X^+)^T w)^T XX^+\alpha$$
$$+ \frac{1}{2}\|y - (X^+)^T w\|_2^2$$
$$+ (X^+y)^T w - \frac{1}{2}\|(X^+)^T w\|_2^2$$

$$= \min_{\alpha} \frac{1}{2}\|XX^+\alpha - (y-(X^+)^T w)\|_2^2$$
$$+ (X^+y)^T w - \frac{1}{2}\|(X^+)^T w\|_2^2$$

And now we can change the optimization back!

as $XX^+\alpha = X\beta$ for some $\beta \in \mathbb{R}^p$

but then we see $\alpha^* = (y - (X^+)^T w)$

Then using the identity $XX^+(X^+)^T = (X^+)^T$,

$$\|XX^+\alpha^* - (y - (X^+)^T w)\|_2^2$$

$$= \|XX^+_y - y\|_2^2$$

which simplifies the original problem to

$$\max_w \; (X^+ y)^T w - \|(X^+)^T w\|_2^2 + \frac{1}{2}\|(I - XX^+)y\|_2^2$$

$$\text{st} \quad \|w\|_\infty \le \lambda$$

$$w \in \operatorname{span} X^T$$

With full-rank columns, i.e. rank $X = p$, $X^+ = (X^T X)^{-1} X^T$ and $w \in \operatorname{span} X^T$ holds, so we may simplify

$$\max_{w} \; \beta_{LS}^{T} w + \frac{w^{T}(X^{T}X)^{-1}w}{2} + \frac{y^{T}(I-P)y}{2}$$

$$\text{s.t.} \quad \|w\|_{\infty} \leq \lambda$$

where $\beta_{LS} = (X^{T}X)^{-1}X^{T}y$

$$P = X(X^{T}X)^{-1}X$$

since $I-P$ is idempotent.

Note $\frac{y^{T}(I-P)y}{2} = \frac{1}{2}\|(I-P)y\|_2^2$

$$= \frac{RSS(\beta_{LS})}{2}$$

Extending to ϕ_i:

generally seems harder, I got an np-harder dual that way, which seems useless.